MPEG-Based Vision

笹　木　美樹男

Last modified on Jan. 5, 2013.
最終更新日：2013 年 1 月 5 日 20 時 40 分

本稿は現在作成中の個人論文であり、今後も改訂されます。

昨今の様々な PC セキュリティ上の懸念に基づき、完成前にハッキング等によって盗用されることのないよう、あえて事前に Web 公開することで筆者のオリジナリティを主張するものです。

（この方針は「神の関数」の HP 全体の著作物に共通します。）

ご理解ください。

1. はじめに
2. 関連研究
3. 理論

3.1 はじめに

本論文で描く理論と実践の概観を図 1 に示す．本論文で提唱する画像認識システムの特徴は以下のとおりである：

●通信、記録、蓄積、編集、インターネット、クラウドと親和性が高い

●特に、MPEG の特徴量をターゲットとして上記の親和性を有する

●認識における記憶・知識処理・予測の重要性に焦点をあて、無記憶条件下での単独の認識器に比べ、複数の認識器の協調や記憶による性能向上を可能にする

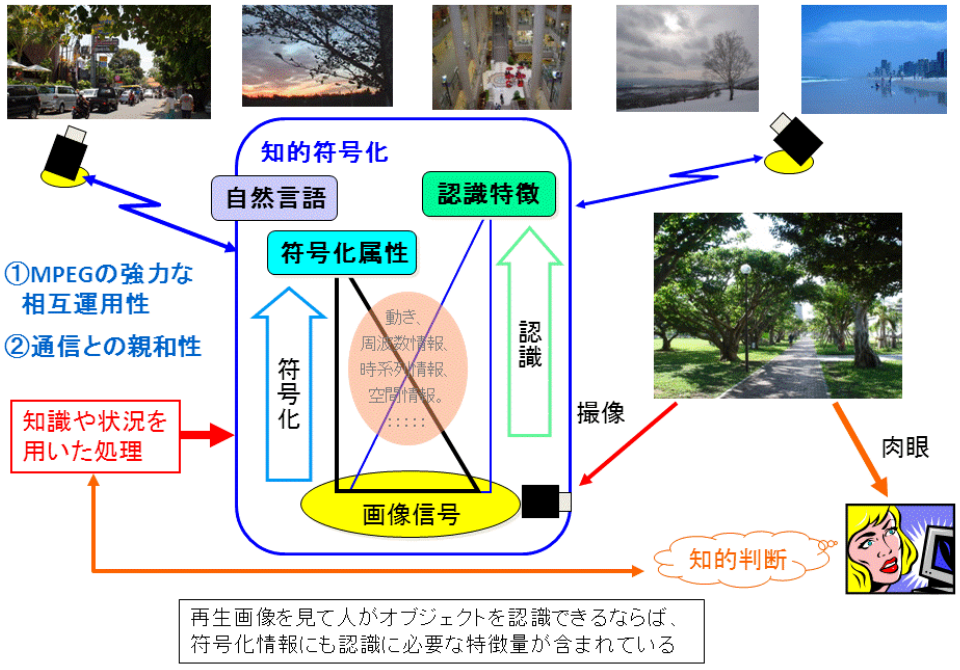


図 3.1 本研究の概観

Fig. 3.1 Targeted overview.

3.2 確信度による認識

3.2.1 確信度ベクトル

各画素ブロックについて、システムは色、テクスチャ、および動き情報をダイレクトに MPEG 符号化データから取り出す。ここで、復号画像の再処理はなくてもよい。そして、ルールベース推論、多変量回帰分析に基づく推論、および簡単な 3 D モデルベースの認識が統合され、各ブロック画素を N 個のオブジェクトカテゴリにマッピングする確信

度ベクトルを算出する。

いま、画像中のある画素領域R（ここでは１つの画素ブロック）に付与するインデックスA_i $(i=1,..,N)$ の確からしさを確信度C_iと定義し、

$$\boldsymbol{C}=\left(C_1,C_2,\ldots,C_i,\ldots,C_N\right)^T \tag{3.1}$$

を確信度ベクトルと呼ぶことにする。

この確信度ベクトルは色やブロック位置，他のオブジェクトとの関係など様々な判定規則で個別に算出した部分的な確信度ベクトルの和として次式で表現する．

$$\boldsymbol{C}=\boldsymbol{c}_0+\sum_{m=1}^{M}\boldsymbol{c}_m \tag{3.2}$$

ただし，$\boldsymbol{c}_0$：初期条件として与えられる確信度ベクトル（N 次元列ベクトル）

$\boldsymbol{c}_m$：m 番目（m=1,...,M）の判定規則に対応する部分的な確信度ベクトル（N 次元列ベクトル）

である．この確信度ベクトルの最終値において，最大成分に相当するインデックスをマクロブロックに該当するオブジェクトであると判断する．

この手法の利点は従来のオブジェクト認識で用いられてきた特徴空間の分割に基づく手法に比べて，考えられる複数の解釈に対応する確信度という形で認識の状態を保持し，随時ある基準の元で唯一の解釈を動的に選択できるところにある．

3.2.2 確信度ベクトルの推定

確信度ベクトルを推定するための基本的手法としてはルールベースで算出する方法と統計的に推論する方法がある．前者は容易に適用できる反面，様々なシーンやオブジェクトに適応させることが困難であり，応用範囲が限定される．一方，後者は適応性が高い半面，学習操作に手間を要する．したがって，ルールベースと統計的推論の両方をうまく融合することがシステム構築の鍵となる．

3.2.3 確信度の線形回帰モデル

確信度ベクトルは大きく分けて２つのクラスの特徴要因に影響される．１つはシーンの状況に関する特徴要因であり，カメラがシーンを撮影した際の時間や場所，天候などがそれに相当する．これについては後述の予測動作（3.18）において再度詳細に取り上げる。もう１つのクラスは画像に関する特徴要因であり，撮影された画像中のあるブロック画素の色，テクスチャ、動き，2次元位置，推定される3次元位置などの画像特徴に関するものがそれに相当する．いま，特徴要因と確信度ベクトルとの間の因果関係を次のような多変量線形回帰モデルで表現する．

$$\boldsymbol{C}=\boldsymbol{c}_0+W\boldsymbol{s} \tag{3.3}$$

ただし，$\boldsymbol{c}_0$は初期確信度ベクトル（N 次元列ベクトル），W は特徴要因 $\boldsymbol{s}$ に対する回帰係数の組を表す N 行 L 列の行列，$\boldsymbol{s}$ は特徴要因を表す L 次元列ベクトルである．

3.3 統計的な推論

3.3.1 多変量回帰分析による学習

複数の解釈結果に対する確信度で構成される空間 C と入力画像や付随情報から得られ

る特徴空間 S との間の回帰関係 W を求める。まず、サンプル映像中の個々のブロック画素の解釈を確信度として教示する。最小２乗法を適用すれば、W^T の推定値 B は次式で計算できる。

$$B=(S^{\mathrm{T}}S)^{-1}S^{\mathrm{T}}Y \tag{3.4}$$

ただし、

$$S=[\boldsymbol{s}(1)\ldots\boldsymbol{s}(k)\ldots\boldsymbol{s}(K)]^T \tag{3.5}$$

であり、$\boldsymbol{s}(k)$は時刻 T_k（$k=1$、…、K）における $\boldsymbol{s}$ の値とする。また、Y は確信度の履歴を表す行列であり、

$$Y=[\boldsymbol{Y}_1\ldots\boldsymbol{Y}_n\ldots\boldsymbol{Y}_N] \tag{3.6}$$

である。ここで、$\boldsymbol{Y}_n$ は n 番目の語彙に対する確信度の時系列を表す K 次元列ベクトルであり、

$$\boldsymbol{Y}_n=\left(y_n(1),\ldots,y_n(k),\ldots,y_n(K)\right)^T \tag{3.7}$$

である。更に、$\boldsymbol{y}(k)$ は時刻 T_k における $\boldsymbol{C}$ の観測値を表す N 次元列ベクトルであり、

$$\boldsymbol{y}(k)=(y_1(k),\ldots y_n(k),\ldots,y_N(k))^T \tag{3.8}$$

である。この観測は本稿では人の判定によって行われる。

式(3.2)と式(3.3)により M 種類の状況要因と確信度ベクトルの間の因果関係を表すと次式のようになる：

$$\boldsymbol{C}=\boldsymbol{c}_0+\sum_{m=1}^{M}W_m\boldsymbol{s}_m \tag{3.9}$$

ただし、$\boldsymbol{c}_0$ は初期確信度を表す N 次元列ベクトルである。W_m $(m=1,\ldots,M)$は要因 $\boldsymbol{s}_m$ に対応する回帰係数行列 $(N\times L_m)$ であり、$\boldsymbol{s}_m$ は L_m 次元列ベクトルを示す.ここで、$i\neq j$ ならば $\boldsymbol{s}_i$ と $\boldsymbol{s}_j$ の間の相関はないものとした.

以下、式(3.3)～(3.6)と同様にして、

$$\boldsymbol{B}_m\equiv\boldsymbol{W}_m{}^T=\left(\boldsymbol{S}_m{}^T\boldsymbol{S}_m\right)^{-1}\boldsymbol{S}_m{}^T\boldsymbol{Y}\qquad(m=1,\ldots,M) \tag{3.10}$$

ただし,

$$S_m=[\boldsymbol{s}_m(1)\ldots\boldsymbol{s}_m(k)\ldots\boldsymbol{s}_m(K)]^T \tag{3.11}$$

であり、$\boldsymbol{s}_m(k)$ は時刻 T_k（$k=1,\ldots,K$）における $\boldsymbol{s}_m$ を示す．また，Y は次式のように，確信度の履歴を表す：

$$Y=[\boldsymbol{Y}_1\ldots\boldsymbol{Y}_n\ldots\boldsymbol{Y}_N] \tag{3.12}$$

ただし，$\boldsymbol{Y}_n$ は K 次元列ベクトルであり，n 番目の語彙に対応する確信度の履歴を表す.

3.4 シーンの認識

以上により、過去の画像の特徴要因と確信度の履歴を学習する（ここでは MPEG-1 形式のブロック画素単位で）と、新たな画像から抽出した特徴要因から確信度ベクトル $\boldsymbol{C}$ を推定するための回帰関係 W を獲得できる。この $\boldsymbol{C}$ の最大成分に対応するインデックスが各ブロック画素について最適なインデックスであると仮定し、それを第１次の認識結果とする。ただし、この処理は後述するターゲットシーンクラスに関して事前に用意した辞書の範囲内で行われる。

3.5 統計的推論手法

時々刻々と移動する車載カメラによる景観認識では次のような性質が望まれる.

① 認識率が改善される

② 認識できるオブジェクトの種類が増える

③ 同じ画素領域に複数の意味を持たせられる

しかしながら，特徴空間をクラスタリングする従来手法では，これらへの対応が困難であった.

3.6 状況ベクトル

時刻 T_k における MPEG 型式の画像中で m_r 行 m_c 列に位置するマクロブロック $MBK（m_r,m_c,k）(m_r=1,..,M_r,\ m_c=1,..,M_c)$が何のオブジェクトであるかを示す確信度ベクトル $\boldsymbol{C}（m_r,m_c,k）$はシーンの状況に関する特徴要因と画像信号に関する特徴要因を説明変数として推定できる．即ち，

$$C(m_r,m_c,k)=W_S\,\boldsymbol{S}_S(m_r,m_c,k)+W_P\,\boldsymbol{S}_P(m_r,m_c,k) \tag{3.13}$$

と表現できる.

ただし，$\boldsymbol{S}_S(m_r,m_c,k)$は $MBK(m_r,m_c,k)$に投影されるシーンの状況を表す L_S 次元列ベクトルであり，W_S は $\boldsymbol{S}_s$ に対する回帰係数の組を表す N 行 L_S 列の行列である.

また，$\boldsymbol{S}_P（m_r,m_c,k）$は $MBK（m_r,m_c,k）$に含まれる画像信号の特徴を表す L_P 次元列ベクトルであり，W_P は $\boldsymbol{S}_p$ に対する回帰係数の組を表す N 行 L_P 列の行列である.

本稿ではこの $\boldsymbol{S}_S$ を状況ベクトルと呼ぶ．いま，状況ベクトルが下記のように構成されるとする.

$$S_S(m_r,m_c,k)=\left[\boldsymbol{S}_{nenv}{}^T\ \boldsymbol{S}_{location}{}^T\ \boldsymbol{S}_{time}{}^T\right]^T \tag{3.14}$$

ここで，$\boldsymbol{S}_{nenv}\ (m_r,m_c,k)$ は $MBK\ (m_r,m_c,k)$ に関するシーンの自然環境情報を表す L_{nenv} 次元列ベクトルである．また，$\boldsymbol{S}_{location}\ (m_r,m_c,k)$ はシーン中の位置情報を表す $L_{location}$ 次元列ベクトルであり，$\boldsymbol{S}_{time}\ (m_r,m_c,k)$ はシーンの時間情報を表す L_{time} 次元列ベクトルである.

状況ベクトルでは一般に代数演算できない特徴量を表現することが求められるため，各状況を言語ラベル（インデックス）で表現して数量化理論 I 類を適用する．例えば，時間帯や天候は次のように表される.

$$\boldsymbol{S}_{time}=(\text{早朝},\text{朝},\cdots,\text{夕方},\text{夜})^T \tag{3.15}$$

$$\boldsymbol{S}_{nenv}=(\text{晴},\text{曇},\text{雨},\cdots,\text{暴風雨})^T \tag{3.16}$$

ただし，インデックス｛早朝，…，晴，…｝は対応する事象が真のとき 1，偽のとき 0 を設定する.

3.7 計測行列の作成

いま，特徴要因行列 S と確信度行列 Y を用いて，

$$\Psi=[S\ Y] \tag{3.17}$$

で表される行列を計測行列と呼ぶことにする．ただし，

$$S=[\boldsymbol{s}(1)\ldots\boldsymbol{s}(k)\ldots\boldsymbol{s}(K)]^T \tag{3.18}$$

であり，$\boldsymbol{s}\ (k)$は時刻 T_k（$k=1,\ldots,K$）における特徴要因ベクトル $\boldsymbol{s}$ の値とする．また，Y は

式(3.6)と同様に確信度の履歴を表す行列である．ここで，Y に含まれる確信度はサンプル映像中の各ブロック画素に投影されるオブジェクトが何であるかを示す正解（Ground Truth）を教示することで獲得される。この正解値は通常、人間の目視判断や高信頼度のセンシング装置で得た判定値、あるいは過去に蓄積された統計的データによって与えられる。このような正解値は，その判定結果であるオブジェクトインデックスに対応する成分の値を 1 とし，それ以外は 0 として設定する．この計測行列から式(3.3)の回帰係数行列を算出すれば，新たな入力映像中のブロック画素に対する確信度ベクトルを推定することができる．いま，

$$\boldsymbol{s} = (\ \boldsymbol{S}_P{}^T\ \boldsymbol{S}_S{}^T\ )^T \tag{3.19}$$

と定義する．ただし，

$$\boldsymbol{S}_p = \left(\boldsymbol{S}_{color}{}^T\, \boldsymbol{S}_{texture}{}^T\, \boldsymbol{S}_{block\_position}{}^T\, \boldsymbol{S}_{motion}{}^T\right)^T \tag{3.20}$$

$$\boldsymbol{S}_{color} = \left(I_Y, I_U, I_V\right)^T \tag{3.21}$$

$$\boldsymbol{S}_{texture} = \left(A_{hor}, A_{ver}, A_{all}\right)^T \tag{3.22}$$

$$\boldsymbol{S}_{block\_position} = \left(m_c, m_r\right)^T \tag{3.23}$$

$$\boldsymbol{S}_{motion} = \left(v_x, v_y\right)^T \tag{3.24}$$

であり，$\boldsymbol{S}_{color}$ は $MBK\ (m_r, m_c, k)$ に関するシーンの色情報を表す L_{color} 次元列ベクトルである．式(3.21)はブロック内平均色であり，2 次元 DCT 係数の DC 成分に対応する．$\boldsymbol{S}_{texture}$ は BLK の 2 次元 DCT 係数の AC 成分から算出する特徴ベクトルであり，A_{hor} は水平方向の画素値変化に対応した AC 成分の絶対値和，A_{ver} は垂直方向の画素値変化に対応した AC 成分の絶対値和，A_{all} はブロック全体にわたる画素値変化に対応した成分の絶対値和、即ち AC 電力を表す．また，$\boldsymbol{S}_{motion}$ は $MBK\ (m_r, m_c, k)$ における 2 次元の動きベクトル（画素単位）を表す．

3.8 認識率

上記で得られた確信度ベクトルの最大成分を求めると確信度が最大であるオブジェクトのインデックス $X\ (m_r, m_c, k)$ を決定できる．これを予め求めておいた正解 $G_{view}\ (m_r, m_c, k)$ と比較して一致するなら $MBK\ (m_r, m_c, k)$ の認識結果 $R_{view}\ (m_r, m_c, k)$ を 1，一致しないならば 0 とする．ただし，$G_{view}\ (m_r, m_c, k)$ は，$MBK\ (m_r, m_c, k)$ の Ground Truth が視点 $view$ の対象とするオブジェクトインデックス群（以下、オブジェクトリスト）O_{view} に含まれるオブジェクトインデックスのどれかに等しいとき 1 を取る．それ以外は認識対象としないため 0 とする．

これに基づき、ある応用上の視点 view に基づいて算出する認識率 Q_{view}（%）を次式で定義することができる．

$$Q_{view} = \frac{100}{K_{view}} \sum_{i=1}^{Mr} \sum_{j=1}^{Mc} R_{view}(i, j, k) \tag{3.25}$$

ただし，

$$K_{view} = \sum_{i=1}^{Mr} \sum_{j=1}^{Mc} G_{view}(i, j, k) \tag{3.26}$$

である．

たとえば安全性という視点 *safe* から見た認識率 Q_{safe} はオブジェクトリスト O_{safe}=｛先行車，対向車，信号，標識｝に基づいて計算することができる．同様に利便性という観点 *convenience* から見た認識率 $Q_{conenience}$ はオブジェクトリスト O_{conv}=｛ビル，緑地，空，先行車，対向車，信号｝に基づいて計算できる．

一方、観光という視点 *sightseeing* から見た認識率 $Q_{sightseeing}$ はオブジェクトリスト $O_{sightseeing}$=｛空、緑地、ビル、海、砂浜｝に基づいて計算できる。ただし、*view* やインデックスの語彙はアプリケーションやユーザプロファイルに依存して決まる。

式(3.25)は画面全体にわたってある視点 view に対応するオブジェクトリスト中に含まれるすべてのオブジェクトについての認識率を評価する尺度であるため，シーン認識率と呼ぶことにする．

これに対し，個々のオブジェクトに対する認識性能を評価する際には、情報検索効率の評価で用いる F 尺度（適合率と再現率の調和平均）をブロック画素単位で算出し，その値をオブジェクト認識率と呼ぶことにする[A12].

3.9 推論特性の合成

多変量線形回帰分析では教示データの追加更新に際して，回帰係数行列を逐次更新式で算出できるという計算上の利点を有する．これは少量の教示データが時々刻々と追加される際の応用に適している．加えて，1 台あるいは複数の車両の車載カメラ映像で得た複数の教示データや推論特性を適宜合成してよりよい推論特性を得ることができれば，冒頭で紹介した数々の応用の実現に大きく貢献できる．

この際に教示データの段階で合成して再度学習させることがもっとも平易であるが，一方で行列の計算量が大きくなりすぎるという問題が発生する．そこで，過去の各学習で既に獲得した特性を再利用することを考えた．これは多変量線形回帰分析の線形性に着目して下記のように導出される．

3.9.1 回帰係数行列の合成

ある教示データセット T が確定していれば，式(3.10)に基づいて m 番目の要因に関する回帰係数行列 B_m が計算できる．いま，議論を簡単にするために 1 つの要因に固定する．そして添え字 m を要因番号から教示データセットの番号 h に読み替えると，

$$\boldsymbol{B}_h = (S_h{}^T\ S_h)^{-1}\ S_h{}^T\ Y_h = (A_h)^{-1} S_h{}^T\ Y_h \tag{3.27}$$

ただし，

$$A_h = S_h{}^{\mathrm{T}}\ S_h \quad (h=1,\ldots,H) \tag{3.28}$$

である．このとき，H 個のデータセット全体を用いて計算される回帰係数行列 $B_{1:H}$ は次式で表される．

$$B_{1:H} = \left\{\sum_{h=1}^{H} A_h\right\}^{-1} \left\{\sum_{h=1}^{H} A_h B_h\right\} \tag{3.29}$$

したがって，個々のデータセットに関して共分散行列 A_h と回帰係数行列 B_h が各々の学習段階で個別に計算されていれば，それらを用いて上式で $B_{1:H}$ を合成できる．

3.9.2 計算量の評価

いま，式(3.29)の計算量を Q_1，式(3.10)の計算量を Q_2 とすると，次式で表現される．

$$Q_g = Q_{g\_inv} + Q_{g\_mul} + Q_{g\_add} \quad (g=1, 2) \tag{3.30}$$

ただし，Q_{g_inv} は次数 $L \times L$ の逆行列計算，Q_{g_mul} は乗算数，Q_{g_add} は加算数である．L 次元正方行列の逆行列計算は式(3.29)，式(3.10)ともに 1 個を要するので，$Q_{1\_inv}=Q_{2\_inv}$ となる．そこで，Q_{g_mul} に着目し，乗算回数の比 α について下限値を評価すると，

$$\alpha = \frac{Q_{2\_mul}}{Q_{1\_mul}} > \frac{2K_0}{N} \tag{3.31}$$

となる．ただし，N は語彙数である．また，簡単化のため，K_0 を 1 個の教示データに含まれる教示イベント数の平均値と考え，$K=HK_0$ とした．一例として N=64，K_0=1000 とすると乗算個数は 1/30 以下にできることがわかる．さらに，式(3.29)の方法は計算に必要なデータのサイズが K_0 に影響されないため，データ通信や記憶において式(3.10)よりも有利である．

3.10 主成分分析による多変量回帰分析

まず、特徴ベクトルをMPEG形式におけるM_{walk} 個の連続する予測フレームから生成する。ここで、n番目の特徴ベクトル$(n=1,\ldots,N)$は次式で定義される。

$$\boldsymbol{x}_n = \left(x_{n1}, \cdots, x_{nP}\right)^T \tag{3.32}$$

ここで、P は特徴ベクトルの次元数である。7 章の実験では仮に M_{walk} を 5 に設定したので、1 秒に 10 フレームの予測フレームがある場合、時系列の区間は 0.5 秒である。したがって N は 64×5=320 となる。

対応する計測行列は次式で表現される。

$$X = \left(\boldsymbol{x}_1 \cdots \boldsymbol{x}_N\right)^T \tag{3.33}$$

この行列は、人間が映っている画像の特性を効果的に学習するために、歩行者シーンから抽出したデータをもとにして定義する。したがって、共分散行列 S は次式で定義される：

$$S = \frac{1}{N} X^T X \tag{3.34}$$

S の p 番目の固有ベクトル$(p=1,\ldots,P)$は次式で記述される。

$$\boldsymbol{w}_p = \left(w_{p1}, \cdots, w_{pP}\right)^T \tag{3.35}$$

主成分得点は次式を用いて計算できる：

$$\boldsymbol{f}_p = X\boldsymbol{w}_p \tag{3.36}$$

これは容易に正規化行列形式に拡張され、次式で表される：

$$F = \left(\boldsymbol{f}_1 \cdots \boldsymbol{f}_P\right) = \mathrm{XWdiag}(1/\sqrt{\lambda_1} \cdots 1/\sqrt{\lambda_P}) = XW\Delta_S^{-1/2} \tag{3.37}$$

ただし、$W = \left(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_P\right)$ であり、λ_p は S の p 番目の固有値を表す。すなわち $\Delta_S \equiv diag\left(\lambda_1, \cdots, \lambda_P\right)$である。

あるシーン中でカテゴリ‘cat’に対応する時区間についてPCAを施し，固有値行列A_{cat}が獲得される場合を考える。いま、人間による教示でシーン中の歩行者（カテゴリ名’ped’で表す）を規定すれば、歩行者がシーン中で映っている時区間における主成分得点（以後、PCSとする）は次式で算出できる：

$$\varphi_{ped} = \boldsymbol{A}_{ped}\boldsymbol{x} \tag{3.38}$$

同様にして、「運転（車載カメラが搭載された車が動いている）中で歩行者がカメラ視野内に存在しない」時区間のカテゴリを‘$drive$’と定義すれば，‘$drive$’に対応する PCS は次

式で表される：

$$\varphi_{drive} = \boldsymbol{A}_{drive}\boldsymbol{x} \tag{3.39}$$

このように，PCA の学習段階で学習データをカテゴリごとに集めることにより，そのカテゴリの事象を認識する際の判断基準となる PCS を算出する固有値行列を獲得できる．

3.11 カーネル主成分分析

KPCA (Kernel PCA) は信号処理分野において、従来のPCAよりも高度な品質を達成するために導入されている。KPCAは入力信号の原空間を高次元空間に射影することで線形識別では困難であった識別を容易にする。線形主成分分析（LPCA）と比較すると、射影後の空間は主軸が湾曲している非線形多様体と考えることができる。以下、文献[G3]を参考にしてKPCAの学習フェーズと認識フェーズの算出方法を示す。

3.11.1 学習フェーズ

原空間のベクトル $\boldsymbol{x}$ で構成される学習データを学習する際に、$\boldsymbol{x}$ を高次空間へ写像する非線形関数を ϕ とすれば、カーネル行列 K の i 行 j 列成分は次式で表される。

$$K_{ij} = \boldsymbol{\Phi}(\boldsymbol{x}_i)\cdot\boldsymbol{\Phi}(\boldsymbol{x}_j) \tag{3.40}$$

これを正規化したカーネル行列は次式で与えられる。

$$\begin{aligned}\overline{K}_{ij} &= \overline{\boldsymbol{\Phi}}(\boldsymbol{x}_i)\cdot\overline{\boldsymbol{\Phi}}(\boldsymbol{x}_j)\\ \overline{K} &= K - 1'_N K - K1_N + 1'_N K 1_N\end{aligned} \tag{3.41}$$

以下、簡単に証明しておく。

$$\begin{aligned}\overline{K}_{ij} &= \overline{\boldsymbol{\Phi}}(\boldsymbol{x}_i)\cdot\overline{\boldsymbol{\Phi}}(\boldsymbol{x}_j) = \left\{\boldsymbol{\Phi}(\boldsymbol{x}_i) - \frac{1}{N}\sum_{m=1}^{N}\boldsymbol{\Phi}(\boldsymbol{x}_m)\right\}\left\{\boldsymbol{\Phi}(\boldsymbol{x}_j) - \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\Phi}(\boldsymbol{x}_n)\right\}\\
&= \boldsymbol{\Phi}(\boldsymbol{x}_i)\cdot\boldsymbol{\Phi}(\boldsymbol{x}_j) - \left\{\frac{1}{N}\sum_{m=1}^{N}\boldsymbol{\Phi}(\boldsymbol{x}_m)\right\}\overline{\boldsymbol{\Phi}}(\boldsymbol{x}_j) - \overline{\boldsymbol{\Phi}}(\boldsymbol{x}_i)\left\{\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\Phi}(\boldsymbol{x}_n)\right\} + \frac{1}{N^2}\left\{\sum_{m=1}^{N}\boldsymbol{\Phi}(\boldsymbol{x}_m)\right\}\left\{\sum_{n=1}^{N}\boldsymbol{\Phi}(\boldsymbol{x}_n)\right\}\\
&= K_{ij} - \frac{1}{N}\sum_{m=1}^{N}\{\boldsymbol{\Phi}(\boldsymbol{x}_m)\boldsymbol{\Phi}(\boldsymbol{x}_j)\} - \frac{1}{N}\sum_{n=1}^{N}\{\boldsymbol{\Phi}(\boldsymbol{x}_i)\boldsymbol{\Phi}(\boldsymbol{x}_n)\} + \frac{1}{N^2}\sum_{m=1}^{N}\left\{\boldsymbol{\Phi}(\boldsymbol{x}_m)\sum_{n=1}^{N}\boldsymbol{\Phi}(\boldsymbol{x}_n)\right\}\\
&= K_{ij} - \frac{1}{N}\sum_{m=1}^{N}\{\boldsymbol{\Phi}(\boldsymbol{x}_m)\boldsymbol{\Phi}(\boldsymbol{x}_j)\} - \frac{1}{N}\sum_{n=1}^{N}\{\boldsymbol{\Phi}(\boldsymbol{x}_i)\boldsymbol{\Phi}(\boldsymbol{x}_n)\} + \frac{1}{N^2}\sum_{m=1}^{N}\left\{\sum_{n=1}^{N}\boldsymbol{\Phi}(\boldsymbol{x}_m)\boldsymbol{\Phi}(\boldsymbol{x}_n)\right\}\end{aligned} \tag{3.42}$$

この正規化カーネル行列を用いて固有ベクトルとその正規化の計算は以下に定義される。

$$\begin{aligned}\hat{\lambda}\boldsymbol{\alpha} &= \overline{K}\boldsymbol{\alpha}\\ \hat{\boldsymbol{\alpha}}^{(l)} &= \frac{\boldsymbol{\alpha}^{(l)}}{\sqrt{\hat{\lambda}_l}} \qquad l = 1,\ldots,N\end{aligned} \tag{3.43}$$

3.11.2 認識フェーズ

認識フェーズでは新しい入力ベクトル $\boldsymbol{y}$ と学習ベクトル $\boldsymbol{x}$ の間で次式によりカーネル行列を計算する。

$$K_{lj}^{in} = \boldsymbol{\Phi}(\boldsymbol{y}_l)\cdot\boldsymbol{\Phi}(\boldsymbol{x}_j) \tag{3.44}$$

式(3.41)と同様に正規化カーネル行列について次式の関係が成立する。

$$\overline{K}^{in} = K^{in} - 1'_N K - K^{in}1_N + 1'_N K 1_N \tag{3.45}$$

高次元空間での主成分を抽出するために固有ベクトル $\boldsymbol{v}$ 上への写像を考えると、

$$\left(\boldsymbol{v}^{(l)} \cdot \overline{\boldsymbol{\Phi}}(\boldsymbol{y})\right) = \sum_{i=1}^{N} \hat{\alpha}_i^{(l)} \left(\overline{\boldsymbol{\Phi}}(\boldsymbol{x}_i) \cdot \overline{\boldsymbol{\Phi}}(\boldsymbol{y})\right) = \sum_{i=1}^{N} \hat{\alpha}_i^{(l)} \overline{K}^{in}(\boldsymbol{x}_i \cdot \boldsymbol{y}) \tag{3.46}$$

となるので、式(3.45)により高次元空間での主軸 $v^{(l)}$に対する射影値（主成分得点）を各 l について算出できる。この N 個の主成分得点をもとに入力ベクトルと学習ベクトルの間の類似性（例えば学習データが歩行者であれば歩行者らしさ）に関する評価値を算出することができる。

3.12 シーンの構造モデル

一般に，走行車両は局所的には前方に直進することが多いため，車載カメラによる道路画像では図 3.2 のような箱型の立体構造モデルを仮定することが考えられる[A8][A9]。

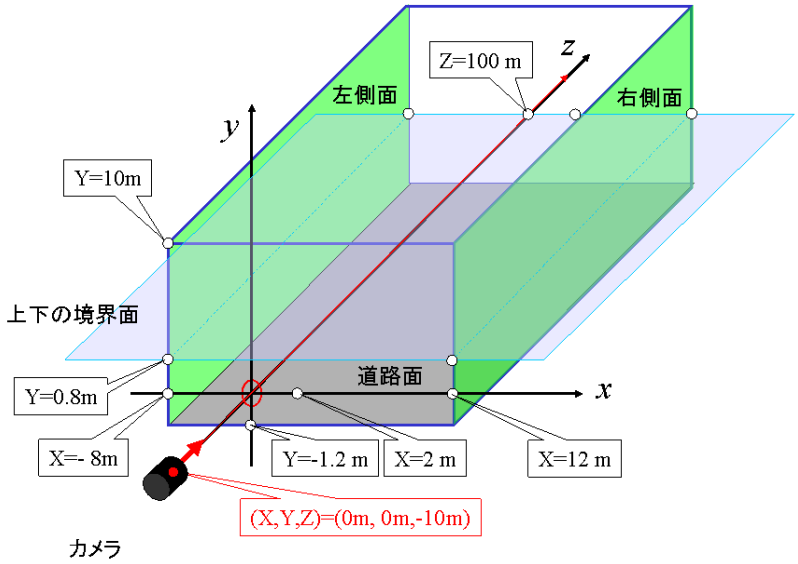


図 3.2 道路シーンの立体構造モデル

Fig. 3.2 Solid structure model of road scene.

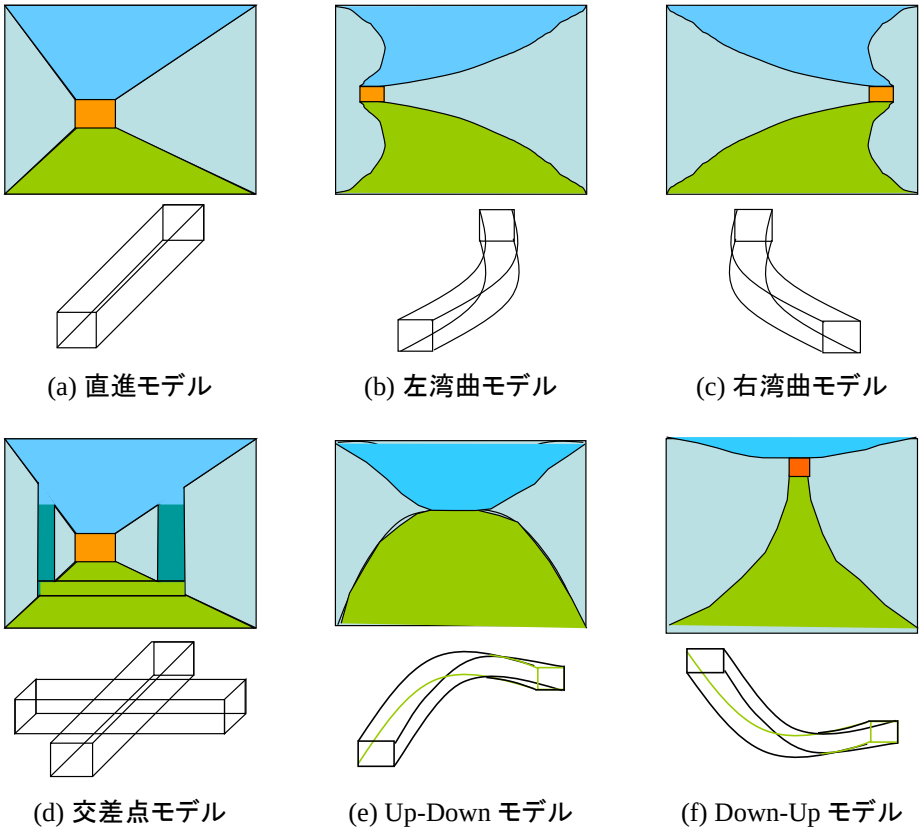


図 3.3 道路シーンの立体構造モデル

Fig. 3.3 Solid structure model of road scene.

3.13 シーンクラス

3.13.1 主観的シーンクラス

主観的シーンクラス(SSC)は人間の定義したシーンの主観的なグルーピングによって決められる。それは画像に含まれる地理的な条件やオブジェクトを用いることで行われる。最初に、筆者は以下の統計的要因に注目することでシーンクラスを考えた。

(1) シーン中の主要なオブジェクトの配置（ビル、樹木、緑地、高架、雪、他）

(2) 道路クラス（一般道、高速道路、他）

(3) 地理的分類（山、市街地、海岸、湖畔、他）

これらの要因の経験的考察によって、図 3.4 に示すようないくつかのクラスが考えられる。それらの各々を自然言語表現することは比較的単純である。実際のシーンでは車両や歩行者といった動的な物体が付加的に存在する。さらに、天候や照明の変化、カメラの位置や姿勢の違い、車両からの見え方（aspect）といった他の要因群も存在する。

3.13.2 計算的シーンクラス

計算的シーンクラス（CSC: Computational Scene Class）は複数のシーンを純粋に機械学習に従って分類しようとするものである。今、

「もし S_1 がシステムに学習されているならば、システムは S_2 を認識できる」

という概念を有向枝で定義できる。これは、より具体的に表現すると、例えば、

「シーン S_1 をシステムの学習データとするとき、システムは S_2 を事前に設定した閾値 α よりも高い認識率で認識する」

と書くことができる。

S_2 から S_1 に向けての有向枝と同時に、もし同様な枝が S_2 から S_1 に向けて定義できるならば、その関係は

「S_1 と S_2 が同じ計算的シーンクラスに属する」

と表現される。ただし、これはシーン S_1 が一意的な CSC を持つことを意味しているわけではない。すなわち、複数の CSC を S_1 について定義することが可能である。他方、ある S_1 に対して CSC と同時に存在する 1 つ以上の主観的シーンクラス（SSC）が分かっているならば、SSC と CSC の関係は自動的に構築できる。

3.14 意味との相互作用

ここでは計測と意味の両面から自動インデキシングの統一的手法を提案する。シーンの 3 次元復元においてよく知られた因子分解法は特異値分解（SVD）から導出され、次式で表現される。

$$\tilde{\boldsymbol{W}} = \boldsymbol{MX} \tag{3.47}$$

ここで、$\tilde{\boldsymbol{W}}$ は計測行列$(2F\times P)$ を表し、F 個のカメラから撮影した対象物体の P 個の 3D サンプル点について平均値分離した 2 次元位置を要素として持つ。$\boldsymbol{M}$ $(2F\times 3)$ は正射影の仮定のもとでカメラ特性を表す。そして $\boldsymbol{X}$ $(3\times P)$ は対象物の 3 次元形状を表す。

次に、意味の獲得は多変量線形回帰モデルを用いて次式で表現される。

$$\boldsymbol{C} = \boldsymbol{RS} \tag{3.48}$$

ここで、$\boldsymbol{C}$ は確信度行列$(N\times P)$であり、N はオブジェクトカテゴリの数を意味する。いくつかの応用では、各カテゴリは自然言語の語彙で表現できる。また、それは確信度ベクトルの意味的なマッピングの特質により、PCA における主成分に関係づけることがで

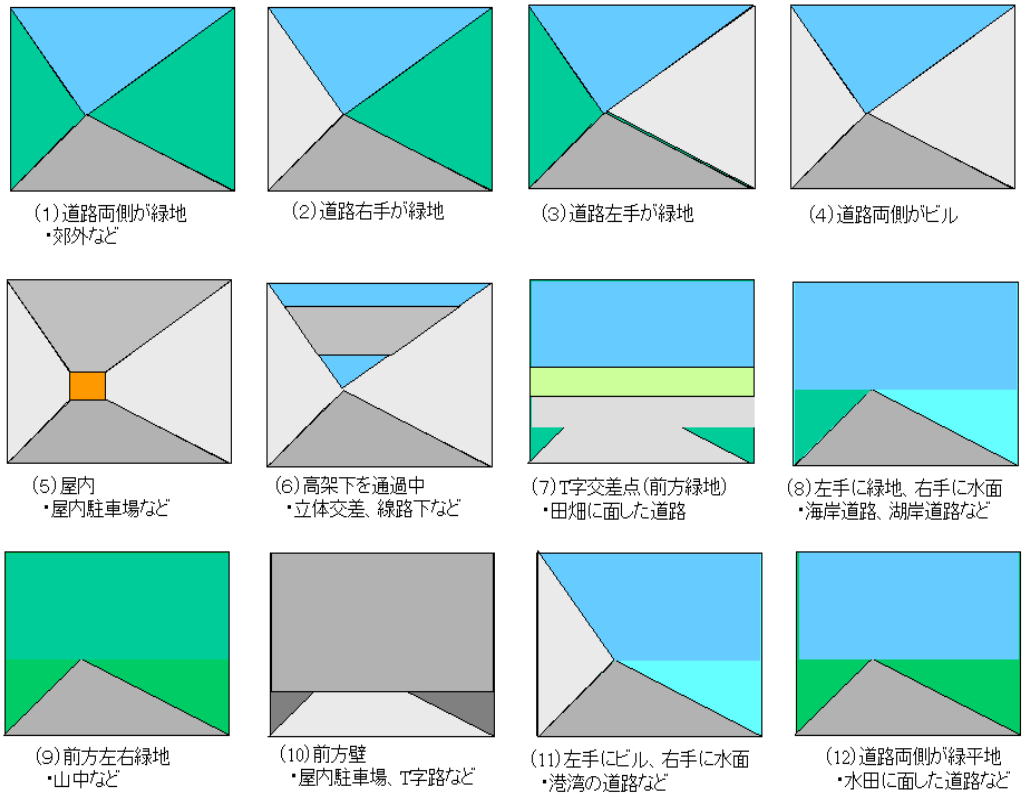


図 3.4 道路シーンの立体構造モデル

Fig. 3.4 Solid structure model of road scene.

きる。それは非線形の場合（カーネル PCA）であっても同様である。$\boldsymbol{R}$ $(N\times L)$ は回帰係数行列を表す。$\boldsymbol{S}$ $(L\times P)$ は観測した状態行列であり、L はターゲットオブジェクト上の各サンプル点における観測状態の数である。基本的にはこの状態はブロック画素に投影された画像情報を通じて獲得される。しかし、容易に周囲の環境や信号やセマンティクスから得られる関連属性に拡張できる。同様に、その拡張により、ユーザプロファイルを含めることも可能である。それは、時間変化特性を有するプロファイルに対しても可能である。分解と同じ形式をとるために、方程式を以下のように変換できる：

$$\boldsymbol{S}=\boldsymbol{R}^{-}\boldsymbol{C}=\boldsymbol{Q}\boldsymbol{C} \tag{3.49}$$

ここで、 $\boldsymbol{Q}$ $(L\times N)$と $\boldsymbol{R}^{-}$は $\boldsymbol{R}$ の一般化疑似逆行列を表す。したがって、$\boldsymbol{S}$ の学習データと $\boldsymbol{C}$ の計測データから、次式の最小２乗法により $\boldsymbol{Q}$ を推定することができる：

$$\boldsymbol{Q}=\boldsymbol{S}\boldsymbol{C}^{T}\left(\boldsymbol{C}\boldsymbol{C}^{T}\right)^{-1} \tag{3.50}$$

式(3.49)はカメラ番号 f $(f=1,\ldots,F)$とは独立であることに注意されたい。したがって、２つの式は次のように統合できる：

$$\overline{\boldsymbol{W}}\equiv\begin{bmatrix}\tilde{\boldsymbol{W}}\\ \tilde{\boldsymbol{S}}\end{bmatrix},\quad \overline{\boldsymbol{W}}=\overline{\boldsymbol{M}\boldsymbol{X}}=\begin{bmatrix}\boldsymbol{M} & \boldsymbol{A}\\ \boldsymbol{B} & \tilde{\boldsymbol{Q}}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{C}\end{bmatrix} \tag{3.51}$$

ただし，$\overline{\boldsymbol{W}}$ $((2+L)F\times P)$ は拡張した計測行列を示す。そして、当面は $\boldsymbol{A}=\boldsymbol{0}$ かつ $\boldsymbol{B}=\boldsymbol{0}$.と仮定する。$\tilde{\boldsymbol{S}}$ $(LF\times P)$と $\tilde{\boldsymbol{Q}}$ $(LF\times N)$はそれぞれ F $\boldsymbol{S}$s と F $\boldsymbol{Q}$s, から容易に作成できる。この方式によりシステムは、膨大な２次元画像メディアデータベースから、対象物とそれを含むシーン全体に関して、その３次元形状と動きのみならずその意味までも同時に獲得

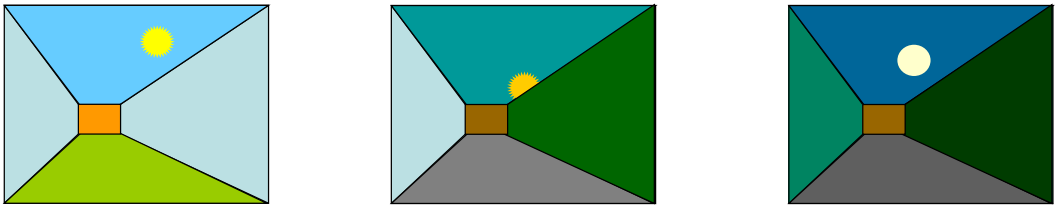
することが可能になる。

しかしながら、式(3.51)にかかわる基本的問題が残されている。一つ目はSVDと多変量回帰（MVR）の違いである。SVDは形状とカメラ運動を同時に導出し、その解は一意的と考えられる。他方、MVRの関係は基本的には射影でありQとCについては複数の解が存在する。２つめの違いはSVDが学習プロセスを要さず、MVRは学習が必要である点である。もし式(3.51)のMVR部分の$\boldsymbol{Q}$がSVD部分の$\boldsymbol{M}$（光学的カメラ特性）と同様の確固たる性質を持っていれば、$\boldsymbol{Q}$と$\boldsymbol{C}$についてもSVDで同時に獲得できる。

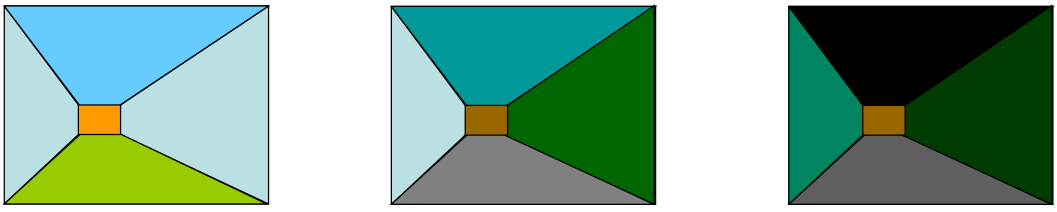
第2のアイデアは式(3.51)を単純に線形射影問題として考えることである。この場合、たとえカメラ特性$\boldsymbol{M}$が透視変換特性を持っているとしても、最小2乗法を使って説くことができる。Mが透視変換特性の場合はSVDでは解けない。これは因子分解法が一般的に正射影を仮定しているためである。このアイデアは式(3.51)が統計的学習問題として考えられているということである。しかしながら、それはカメラ番号とカメラ配置の自由度という問題を引き起こすかもしれない。もしF個のカメラが固定されているかあるいはかなり制約のある軌道上に位置しているとき、この統計的学習は可能になる。

知識処理の観点でいうと、式(3.51)における行列Aと行列Bは特殊な役割を持っている。$\boldsymbol{A}$ は$\boldsymbol{C}$ から $\boldsymbol{W}$,への写像を意味する。すなわち、サンプル点の２次元位置情報は各点が所属するオブジェクトカテゴリの確信度に影響されるかもしれない。第2に、$\boldsymbol{B}$は$\boldsymbol{X}$から$\boldsymbol{S}$ への写像を意味しており、これによって、形状情報が観測状態に影響を与える可能性があることを示している。

さらに、これらの$\boldsymbol{A}$ と $\boldsymbol{B}$ は時空間的な発展過程を持つと同時に、人間のフィーリングや情動にかかわるセンシング特性に関するいくつかの数量的表現への非線形写像を有するであろう。この期待は式(3.51)における基本的な数学的メカニズムが線形システムを志



（1a）昼間（太陽が視界にある）　（2a）夕刻（太陽が視界にある）　（3a）夜間（月が視界にある）



（1b）昼間（太陽が視界にない）　（2b）夕刻（太陽が視界にない）　（3b）夜間（月が視界にある）

図 3.5 道路シーンの立体構造モデル

Fig. 3.5Solid structure model of road scene.

向することを意味する。そしてシステムにおける他の非線形機能は主として非線形マッピング、論理操作、および状態遷移構造のような巡回アーキテクチャで実現される。

観点のカテゴリでいえば、式(3.51)はさらに拡張できる。拡張した観点が$\boldsymbol{U}$ (user),であるとすると、上記ですでに議論した２つの観点は$\boldsymbol{X}$ (3D シーン構造)と$\boldsymbol{C}$ (コンテンツの意味)である。これより式(3.51)は次式のように拡張できる：

$$\boldsymbol{W}^* = \boldsymbol{M}^* \boldsymbol{X}^* \tag{3.52}$$

ただし,

$$\boldsymbol{W}^* \equiv \begin{bmatrix} \boldsymbol{W} \\ \boldsymbol{S} \\ \boldsymbol{P} \end{bmatrix}, \quad \boldsymbol{M}^* \equiv \begin{bmatrix} \boldsymbol{M}_X & \boldsymbol{A}_C & \boldsymbol{A}_U \\ \boldsymbol{B}_X & \boldsymbol{M}_C & \boldsymbol{B}_U \\ \boldsymbol{C}_X & \boldsymbol{C}_C & \boldsymbol{M}_U \end{bmatrix}, \quad \boldsymbol{X}^* \equiv \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{C} \\ \boldsymbol{U} \end{bmatrix} \tag{3.53}$$

である。計測行列 W^*における P は、実際上はたとえば、システムのシーン理解に影響を与えるユーザ属性を数量化理論で表現したユーザプロファイルに相当する。

3.15 アプリケーションと語彙

カメラが車載カメラのように移動するときは、特にそれが前に進み通り抜けるとき、ラベリング問題は突然難しくなる。スナップショットはラベル付けすることができ、シーンクラス（SC）を生成することはできるが、ビデオシーンは絶えず変化し新しい SC を創成する。この結果、たとえ観測したビデオシーンがトータルとして一つのラベルしか持たない場合でも、それは局所的には多くの異なるラベルを含むであろう。

語彙の定義はきわめてターゲットアプリに依存する。シーンクラス獲得で考えた目的は検索、分類、エンタテインメント、ロケーションアウェアサービス、環境認識など多岐にわたる。あるビデオシーケンスについて SC の時間スパンを考えねばならない。これについての簡単なアイデアはあるシーンについての SC ラベルを SC ラベルの時系列の作成を通じて決定していくことである。ここで、シーン中の各単独のフレームについての SC ラベルの各々は次のように定義できる：

$$\begin{aligned} \alpha_k &= Index\left\{ p(\boldsymbol{X}^*_k \middle| \boldsymbol{W}^*_{1:k}, \boldsymbol{X}^*_{1:k-1}) \right\} \\ \alpha &= SceneClass\{\alpha_{1:K}\}, \quad \alpha_{1:K} = \{\alpha_1, \cdots, \alpha_K\} \end{aligned} \tag{3.54}$$

ただし、K はあるシーンに含まれる全フレーム数であり、$p()$ は確率分布を意味する。$Index$ はオブジェクトベースの確信度からフレーム時刻 k における SC ラベルへの写像である。$SceneClass$ は一連のインデックスからシーン全体に割り当てた単一のラベルへの写像である。この式の背後にある基本的なアイデアはパーティクルフィルタリングベースの状態トラッキング[E-7]である。シーンデータベースの定義と集め方はターゲットアプリによって定められ、語彙と調整に用いる論理操作が(3.54)式の２つの非線形写像に含まれる。さらに、この CSC の自動生成機構は非線形写像により時系列表現された特徴のラベリングを可能にするが、カーネル主成分分析（KPCA）や多様体学習[E-12]に関する研究とも関連が深い。

２つ目の問題は視点（viewpoint）の影響である。これはターゲットアプリの制約のもとでの評価側の変化である。後述するように、CSC の有向グラフ表現は視点の影響を受ける。自然言語理解の類推に基づき、単語間の意味ネットワークの有向グラフ構造は、視覚的概念形成に向けた出発点のひとつとして CSC 生成に適用できる。

3.16　複数手法の統合化

3.16.1 3S 認識モデル

この考えは形状、スペクトル、状況という３つの大きなトリガーで有機的に活性化される。まず 1 番目の形状ベースのオブジェクト認識手法は既に数多く存在し、通常はこれがもっともポピュラーなアプローチである。しかしそれらはしばしば、ぼけや隠れ、動的な変化、あいまいさ、個々の差異といった深刻な問題を呈する。

2 番目にスペクトルベースの認識を考える。これは空間スペクトルと時間スペクトルの両方に適用される。このスペクトルベースのアプローチは形状ベースの手法が苦しんできた問題に対してロバストであるが、認識タスク全体を単独でカバーできない。

３番目の状況は、対象領域を信号処理によりパターン分類した中間結果を、知識ベースで判定する際に必要となる。この状況要因は観測者とアプリケーションに依存するが、パーティクルフィルタのような仮説の生成棄却を伴うトラッキングにおいては良好な仮定を提供する。しかし、その際に対象領域の形状とスペクトルが十分に解析されていなければ、必要以上に多くの確率的解釈を生成し、最終解の判定が困難になる恐れがある。

そこで、筆者はこれら３つのトリガーの動的な活性化手法（以下、３S モデル）を提案する。これによってシステムはメインの戦略を随時スイッチングできる。そして確信度は一連のサブタスクを統合し、３S モデルにおけるその動的スイッチング挙動を通じて認識に向かう。

3.17　映像内容の予測と推定

これから撮影しようとするシーンのシーンクラスがシステムに与えられている場合、

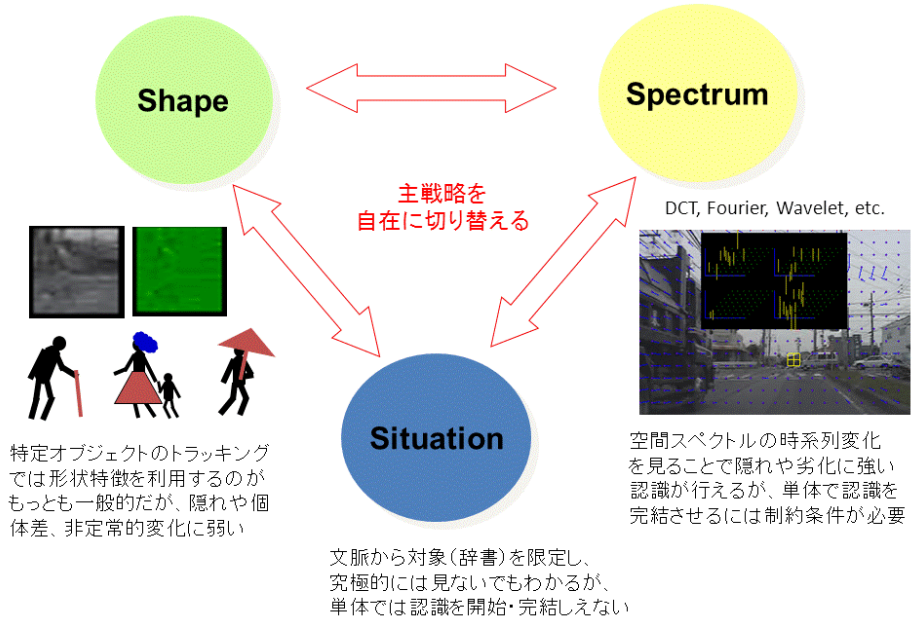


図　3.6 3S モデルの概念

Fig. 3.6　Concept of 3S model.

システムはシーン内のオブジェクトやシーン構造の認識を行うための語彙と認識特性を

事前に用意することができ、解空間の不要な探索はなくなる。これは近年 GPS 機能を有するカメラが増大し、撮影した内容に位置情報を付与できることからきわめて現実的なことである。また、異なる撮影者が Web 上で同一の場所のシーンをアップロードして共有することにより、さまざまな時間条件のもとで過去のシーンが蓄積され、スタティックな事象（建物や地形の形状など）あるいは周期的変化を伴う現象（波が打ち寄せるシーン、木々が揺れるシーン、天候の変化に伴うシーン、四季折々の変化、人の流れなど）については十分予測可能になってきている。その蓄積される画像の付与される情報には次のような場合が考えられる：

（１）場所の情報が与えられている場合

a) カメラの位置姿勢およびその時間変化（軌道と光軸がわかる場合

b) 視野内のオブジェクトやシーン構造が既知または推定できる場合

c) web や地図などにある説明情報（description）が使える場合

（２）時間はわかるが場所の情報が不明の場合

（３）時間も場所も不明の場合

他にも、カメラの各種設定パラメータ（ズームや照度など）が影響することが考えられる。映像の予測と推定にはこのような付帯記述や検索インデックスを有する過去のデータベースが極めて重要と考えられるが、必ずしもそれらは現段階で整備されているわけではない。そこでまず、未整備の過去データを用いる統計的手法を考察する。

いま、次のような時系列モデルを仮定する：

$$\boldsymbol{y}_n = \boldsymbol{t}_n + \boldsymbol{s}_n + \boldsymbol{p}_n + \boldsymbol{d}_n + \boldsymbol{c}_n + \boldsymbol{w}_n \tag{3.55}$$

ここで、$\boldsymbol{y}_n$ は K 次元列ベクトルで、k 番目の成分は時刻 n に k 番目の状態節点（particle）（$k = 1,\ldots,K$）で観測される歩行者の数であり、その状態とは地理的スポットとオブジェクトの組み合わせとして定義される。同様に、$\boldsymbol{t}_n$, $\boldsymbol{s}_n$, $\boldsymbol{p}_n$, $\boldsymbol{y}_n$, $\boldsymbol{c}_n$, $\boldsymbol{w}_n$, はそれぞれトレンド成分、季節成分、定常 AR（自己回帰）成分、曜日成分、時刻の成分、そして白色雑音に対応する K-次元ベクトルである。もし、マクロスケールの状態ベクトル $\boldsymbol{x}_n$ を図 3.7 のように定義すれば、$\boldsymbol{x}_n$ は次式で表現される。

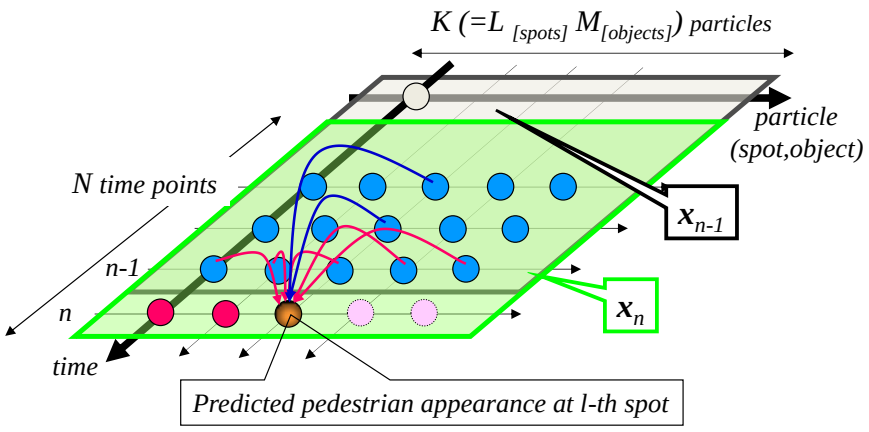


図 3.7 歩行者出現予測のグラフィカルモデル

Fig. 3.7 Graphical model for prediction of pedestrian appearance..

$$\boldsymbol{x}_n = \left(y_{n,1}, \ldots, y_{n,K}, \ldots, y_{n-(N-1),1}, \ldots, y_{n-(N-1),K} \right)^T \tag{3.56}$$

したがって、式(3.55)はよく知られた状態空間モデルに変換され、次式で表現できる：

$$\begin{cases} \boldsymbol{x}_n = F\boldsymbol{x}_{n-1} + G\boldsymbol{v}_n \\ \boldsymbol{y}_n = H\boldsymbol{x}_n \quad + Q\boldsymbol{w}_n \end{cases} \tag{3.57}$$

ただし F, G, H, および Q はそれぞれ、 $KN \times KN$ 行列、$KN \times M_v$ 行列、$K \times KN$ 行列, $K \times M_w$ 行列である。$\boldsymbol{v}_n$ と $\boldsymbol{w}_n$ はそれぞれ、M_v -次元のシステム雑音、M_w-次元の観測雑音である。式(3.55)の右辺における各項を抽出するためには $\boldsymbol{x}_n$ に関していくつかの仮定を追加する必要がある。また、各係数行列の推定に際しては、歩行者数に対応する莫大な量の正解値を用意する必要がある。

3.18 予測動作における状況ベクトル

上記の特徴ベクトル（目的変数）が発生する際に，その説明変数となる状況ベクトルは表 3.2 の｛場所，日時，天候｝で構成される．ここで，場所は 3 次元（各次元が区間に相当），日時は｛年，月，曜日，時間帯｝の合計 25 次元，天候は｛晴，曇，雨｝の 3 次元とした．

状況ベクトルは更に拡張でき，日種効果，季節効果，時間帯効果，年中行事の影響なども考慮することができる．また，道路環境や地理要因も同様に記述できる．本稿では以下の 3 つのベクトルで構成する．

- $\boldsymbol{x}_i$：各成分は番号 i $(i=1,..,I)$を割り振られた地点を地理的に特定する語彙（座標，地名に関する言語表現），すなわち非数値情報に対応し，これらを数量化理論により数値化した値をとる．具体的には，各語彙を予測条件として指定するときに 1，そうでなければ 0 とする．
- $\boldsymbol{t}_j$：各成分は番号 j $(j=1,..,J)$を割り振られた離散時刻を特定する語彙（季節，月，曜日，時間帯）と対応しており，予測条件として指定するときに 1，そうでなければ 0 とする．
- $\boldsymbol{s}_k$：各成分は番号 k $(k=1,..,K)$を割り振られた付帯状況要因を特定する語彙と対応しており，$\boldsymbol{t}_j$ における $\boldsymbol{x}_i$ の付帯状況（天候や周囲環境など）を示す．各要因は予測条件として指定するときに 1，そうでなければ 0 とする．

状況ベクトルは原理的には本来，代数演算可能な数値属性（気温，照度，湿度，道路幅，集客施設までの距離[I2]など）を含むことができる．10 章では記述の容易さを重視し，語彙の数値化情報のみを用いた場合の実験例を示す．

3.19 主成分回帰分析

上記の特徴ベクトルが状況ベクトルに回帰すると仮定して多変量回帰分析を行うと，状況行列が特異となり，共分散行列に対する良好な逆行列を算出できない場合がある．このようなとき，原空間に主成分分析を施して次元を削減すれば各オブジェクトの確信度が予測可能になる．以下，その手順を示す．

まず，上記の$\boldsymbol{x}_i$と$\boldsymbol{s}_j$を成分として構成される状況ベクトルを次式で表す．

$$\begin{aligned}\boldsymbol{S} &= \left(S_1,\cdots,S_L\right)^T = \left[\boldsymbol{S}_{place}{}^T,\boldsymbol{S}_{time}{}^T,\boldsymbol{S}_{env}{}^T\right]^T \\ &= \left[\boldsymbol{x}_i{}^T,\boldsymbol{t}_j{}^T,\boldsymbol{s}_k{}^T\right]^T\end{aligned} \tag{3.58}$$

これにより，{*i*,*j*,*k*}の 3 項組がクラス分けされた状況イベントを構成することがわかる．これを *H* 個のイベントについて収集すると，次式の計測行列 *S* (*H*×*L*)が構成できる．

$$S = \left(\boldsymbol{S}_{[1]}\cdots\boldsymbol{S}_{[H]}\right)^T \tag{3.59}$$

これをもとにして，共分散行列 *V* (*L*×*L*)は次式で計算される．

$$V = \frac{1}{H}S^T S \tag{3.60}$$

この固有値問題を解いて得られる *V* の *p* 番目の固有ベクトル (*p*=*1*,…,*P*) は次式で表現される．

$$\boldsymbol{w}_p = \left(w_{p1},\cdots,w_{pL}\right)^T \tag{3.61}$$

これにより，*p* 番目の主成分に対応する主成分得点を *H* 個の状況ベクトルについて計算した主成分得点ベクトルは次式で表現される．

$$\boldsymbol{q}_p = S\boldsymbol{w}_p \tag{3.62}$$

これを正規化した行列形式 *Q* (*H*×*P*)に拡張して，次式で表す．

$$\begin{aligned}Q &= \left(\boldsymbol{q}_1\cdots\boldsymbol{q}_P\right) \\ &= SWdiag(1/\sqrt{\lambda_1}\cdots 1/\sqrt{\lambda_P}) = SW\Delta_V^{1/2}\end{aligned} \tag{3.63}$$

ただし，$W = \left(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_P\right)$ であり λ_p は *V* の *p* 番目の固有値を表す．すなわち，

$$\Delta_V \equiv diag\left(\lambda_1,\cdots,\lambda_P\right). \tag{3.64}$$

である．

3.20 予測器

ここでは移動物体が出現することを過去の撮影結果から予測することを考える。

3.20.1 基本動作

過去の統計に基く予測頻度の行列 *F*（*H*×*M*）は多変量回帰予測によれば，

$$\begin{aligned}F &= QB \\ Q &= \left[\boldsymbol{Q}^{(1)}\cdots\boldsymbol{Q}^{(H)}\right]^T\end{aligned} \tag{3.65}$$

ただし，*B* は回帰係数行列（*P*×*M*），$\boldsymbol{Q}^{(h)}$は *h* 番目のイベントに対応する主成分空間の状況ベクトル（*P*×*1*），そして *Q* は主成分空間の状況行列（*H*×*P*）を表す．これより *B* は最小 2 乗法を用いて次式で算出できる．

$$B = \left(Q^T Q\right)^{-1} Q^T Y \tag{3.66}$$

ただし，*Y*（*H*×*M*）は *H* 個のイベントにおける *M* 種類のオブジェクト（表 2 では *M*=12）の出現頻度を表すデータ行列である．*Y* は *H* 個の各イベントにおいて画像認識器が出力した確信度行列 *C*（N_B×M_O）から歩行者（表 2 では M_O=*M*/2=6）と判定された画素領域に注目し，*M* 個の出現頻度値に変換することで生成される．ここで，N_B は画像中の認識単位である画素領域の数，M_O は画像認識が判別する歩行者オブジェクトの種類の数とする．いま，この変換操作を次式で表現する．

$$Y = \Psi\left(C\right) \tag{3.67}$$

この C の信頼度は画像認識の認識率 R によって定められるため，C から生成した Y を学習データとする回帰係数行列 B は R の影響を受ける．いま，この不完全な認識器が推定した真理値を Estimated Truth（以下，ET）と呼ぶことにする．一方，人間や完全な認識器が与える正解値は一般に Ground Truth（以下，GT）と呼ばれ，その信頼度はほぼ 100%となる．しかし，ET が機械により自動獲得できるのに対し，GT の獲得は一般に多大なる労力を要する．そこで，ET を学習データとして用いることを想定し，GT も含めて次式で表現する．

$$\hat{Y} = \left\lfloor \hat{f}_{hm} \right\rfloor \quad (h = 1,\ldots,H)(m = 1,\ldots,M) \tag{3.68}$$

ただし，$f\hat{}_{hm}$ は h 番目のイベントに対応する画像中の m 種類目の歩行者オブジェクトの出現頻度を示す．

3.20.2 ET の変化を見積もるモデル

次に，この学習データ $Y\hat{}$ によって獲得できる予測性能を見積もるため，$f\hat{}_{hm}$ を次式でモデル化する．

$$\hat{f}_{hm} = A(r_{hm}, g_{hm}) \tag{3.69}$$

ただし，

r_{hm}：画像認識器の信頼度（平均認識率）．

g_{hm}：出現頻度に関する GT 値．

$A(r,g)$：画像認識で得られる出現頻度の ET 値を認識率 r と GT 値 g から見積もるモデル関数．

である．いま，画像認識率 R を再現率 α と適合率 β の調和平均で定義すると，

$$R = \frac{2\alpha\beta}{\alpha+\beta},\ \alpha = \frac{g-u}{g},\ \beta = \frac{g-u}{g-u+d} \tag{3.70}$$

となる．ただし，g は GT 値，u は未検出個数，d は誤検出個数を表す．このとき，認識率 R で計測される出現頻度 $f\hat{}$ の最大値 $f\hat{}_{max}$ と最小値 $f\hat{}_{min}$ はそれぞれ，

$$\hat{f}_{max} = g\frac{2-R}{R} \geq \hat{f} \geq g\frac{R}{2-R} = \hat{f}_{min} \tag{3.71}$$

と表現される．今，$f\hat{}$ が上記の区間内で同じ出現確率で分布すると仮定すると，その平均値は

$$\begin{aligned} \hat{f}_{avr} &= \frac{\hat{f}_{max} + \hat{f}_{min}}{2} = A_R g \\ A_R &= \frac{1}{R} + \frac{1}{2-R} - 1 \end{aligned} \tag{3.72}$$

となる．したがって，

$$A(r,g) = \left(\frac{1}{r} + \frac{1}{2-r} - 1\right)g \tag{3.73}$$

を式(3.69)に適用すれば $Y\hat{}$ を算出できる．この $Y\hat{}$ を過去の統計データとして予測頻度の行列 F の近似値 $F\hat{}$ を算出すると，

$$\hat{F} = Q\hat{B}$$
$$\hat{B} = \left(Q^T Q\right)^{-1} Q^T \hat{Y} \tag{3.74}$$

となる.

3.20.3 認識器の性能改善

このような予測とそれに対応する観測（画像認識）が過去に H_F 回行われたとすると，予測対象となった地点で車両が実際に観測した出現頻度の行列 F_1 と予測頻度の行列 F_0 との間の誤差⊿Fは次式で表される.

$$\Delta F = F_1 - F_0 \tag{3.75}$$

ただし，

$$F_a = \left[\boldsymbol{f}_{a,1}, \boldsymbol{f}_{a,2}, \ldots, \boldsymbol{f}_{a,H_F}\right]^T \quad (a = 0,1) \tag{3.76}$$

$$F_0 = Q\hat{B} \tag{3.77}$$

である.

車載器，センター，携帯，PC，定点カメラなどによる ET や GT は予測器の学習データとなる．このとき式（3.75）の予測誤差から次式の評価尺度を算出する.

$$E_F = \sum_{i=1}^{H_F}\sum_{j=1}^{M}\left|\Delta F_{ij}\right| \tag{3.78}$$

この E_F がある閾値 E_{FTH} を超えるとき，予測器と認識器を修正すれば半自動的に性能を改善できるであろう．いま，予測対象とする地域に応じて経験的に 2 つの閾値 f_{th} と Φ_{th} を設定し，次の制約式を満たすまで学習データを増やす.

$$\left\{\left\|\boldsymbol{f}(\boldsymbol{S})\right\|_1 \geq f_{th}\right\} AND \left\{\left\|\boldsymbol{\Phi}(\boldsymbol{S},N)\right\|_1 \geq \Phi_{th}\right\} \tag{3.79}$$

ただし，

$$\begin{cases} \boldsymbol{\Phi}(\boldsymbol{S},N) = \left\{\sum_{n=1}^{N} \boldsymbol{r}_n \boldsymbol{f}_n(\boldsymbol{S}) + \boldsymbol{f}_0\right\} \\ \boldsymbol{r}_n = diag\left(r_{n1}, \ldots r_{nm}, \ldots r_{nM}\right) \\ \boldsymbol{f}(\boldsymbol{S}) = \dfrac{1}{(N+1)}\boldsymbol{\Phi}(\boldsymbol{S},N) \\ \boldsymbol{f}_n = \left(f_{n1}, \cdots, f_{nM}\right)^T \end{cases} \tag{3.80}$$

である．ここで，n は認識器の識別番号であり，$n=1,\ldots,N$. $\boldsymbol{f}_0$ は初期の歩行者頻度ベクトル，$\boldsymbol{f}_n$ は認識器 n が出力する歩行者頻度ベクトル，そして，$\boldsymbol{f}$ は N 個の認識器の出力に関する平均値ベクトルである．また，$\boldsymbol{S}$ は認識時の状況ベクトルであり，r_{nm} は認識器 n のオブジェクト m に関する認識率を表す.

この更新動作はセンターから信頼度つきの ET 値または事前に算出した認識特性を配信することで達成される．その後，上述の予測，認識，差分評価のプロセスを繰り返し，E_F が E_{FTH} 以下ならば修正を終了する．$E_F > E_{FTH}$ であれば，認識パラメータを改善し，認識，予測，E_F の最小化を繰り返す.

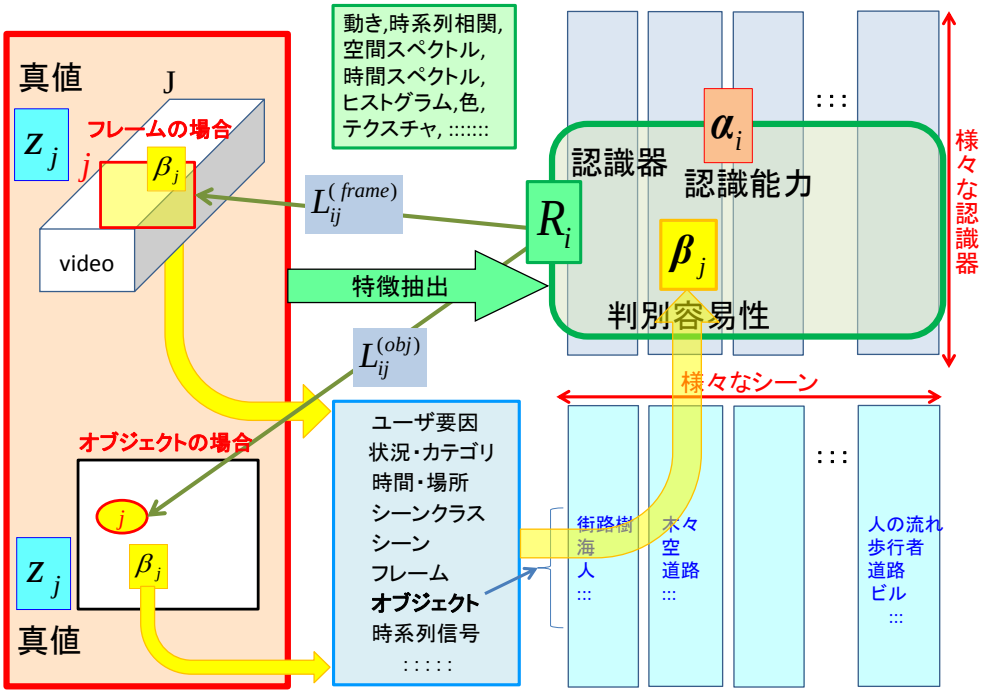


図 3.8 映像と認識器の関係

Fig. 3.8 Relation between images and recognizers.

3.20.4 EM アルゴリズムの応用

複数の認識器を巡回的に協調させるという手法をより高度なものにした例として、EM アルゴリズムを導入したGLAD (Generative model of Labels, Abilities, and Difficulties)手法[H1]がある。この手法を参考にして認識に関連する属性要因を拡張し、MBVに適用する。

いま、K 個のブロック画像 $BLK(k),(k=1,..K)$ の集合を考える。各ブロック画像は N 個の対象カテゴリの 1 つに属するとする。

このとき、ブロック画像 $BLK(k)$ のオブジェクトインデックス z_k を決定したい。観測したインデックスは主としてブロック画像の判別容易性、真のオブジェクトインデックス、認識器の認識能力の 3 つに依存する。

いま、ブロック画像 $BLK(k)$ の判別容易性を次式で表現する：

$$\beta_k \in [0,+\infty) \tag{3.81}$$

また、画像 $BLK(k)$ の真のインデックスを z_k とする。また、認識器 L_i の認識能力のパラメータを次式で表現する：

$$\alpha_i \in (-\infty,+\infty) \tag{3.82}$$

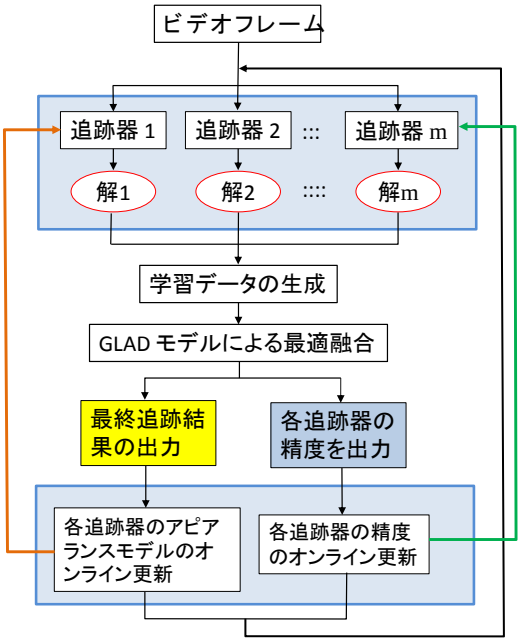


図 3.9 追跡アルゴリズム

Fig. 3.9 Tracking algorithm..

α_i は認識対象とする画像 $BLK(k)$ に関連する以下の要因の各々について定義できる。

Level-7) ユーザ要因 :ユーザが求める認識タスクおよびユーザの嗜好性
Level-6) 状況・カテゴリ :シーン撮影時の状況と分類カテゴリ
Level-5) 時間・場所： :シーンが撮影された時間と場所
Level-4) シーンクラス :シーンクラス（3.13 節参照）
Level_3) シーン
Level_2) フレーム
Level_1) オブジェクト
Level_0) 時系列信号 : $BLK(k)$ の特徴量およびそれを含む時系列

L_i が $BLK(k)$ に与えたインデックスを l_{ik} とし、それが真のインデックスに等しい確率を次式のモデルで表現する。

$$p(l_{ik}=z_k|\alpha_i,\beta_k)=\frac{1}{1+e^{-\alpha_i\beta_k}} \tag{3.83}$$

このモデルのもとで獲得したインデックスで正しいものに関する対数尤度（log odds）は認識能力と判別容易性の双一次関数として次式で表現される。

$$\log\frac{p(l_{ik}=z_k)}{1-p(l_{ik}=z_k)}=\alpha_i\beta_k \tag{3.84}$$

図 3.12 はモデルの因果関係を示すグラフィカルモデルである。 z_k,α_i,β_k は既知の先験的

分布でサンプルされる。これらは式(3.83)によって観測インデックスを決定する。観測イ

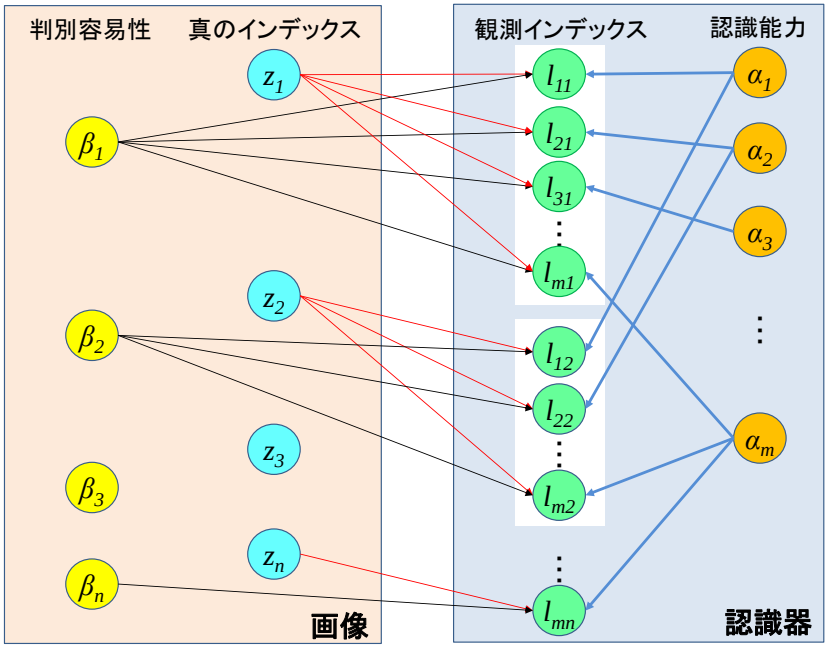


図 3.10 グラフィカルモデル

Fig. 3.10 Graphical Model..

ンデックスの集合 $L=\{l_{ik}\}$が与えられたとき、求められるタスクは $Z=\{z_k\}$, α_i, β_k について最も尤度の高い値を同時に推定することになる。そこで，これらの最尤推定を行うために期待値最大化（EM）手法を適用する：

【E ステップ】

画像 $BLK(k)$について与えられたすべてのインデックスを

$$l_k=\left\{l_{i\hat{k}}\middle|\hat{k}=k\right\} \tag{3.85}$$

と表す。ただし、すべての認識器が各ブロック画像をインデキシングするわけではない。この場合、$l_{i\hat{k}}$中の認識器に関する識別番号 i は画像 $BLK(k)$のインデックスを付与した認識器のみ参照する。すなわち、

$$i\in\{i_L(k,1),\ldots,i_L(k,I_k)\} \tag{3.86}$$

であり、参照した認識器群全体の認識能力を表すベクトル $\boldsymbol{\alpha}(k)$は次式で表される：

$$\boldsymbol{\alpha}(k)=\left(\alpha_{i_L(k,1)},\cdots,\alpha_{i_L(k,I_k)}\right) \tag{3.87}$$

また、K 個のブロック画像全体の判別容易性を表すベクトル $\boldsymbol{\beta}$ は次式で表される：

$$\boldsymbol{\beta}=(\beta(1),\cdots,\beta(K)) \tag{3.88}$$

最新の M ステップにおいて得た $\boldsymbol{\alpha}(k)$, $\boldsymbol{\beta}$ が与えられたときのすべての $z_k\in\{0,1\}$と観測したインデックスについて事後確率は次式で計算する：

$$\begin{aligned}p(z_k|\boldsymbol{l},\boldsymbol{\alpha}(k),\boldsymbol{\beta})&=p(z_k|\boldsymbol{l}_k,\boldsymbol{\alpha}(k),\beta_k)\propto p(z_k|\boldsymbol{\alpha}(k),\beta_k)p(\boldsymbol{l}_k|z_k,\boldsymbol{\alpha}(k),\beta_k)\\&\propto p(z_k)\prod_{i=i_L(k,1)}^{i_L(k,I_k)}p(l_{ik}|z_k,\alpha_i,\beta_k)\end{aligned} \tag{3.89}$$

ここで、 グラフィカルモデルから得られる条件独立仮定により、$p(z_k|\boldsymbol{\alpha},\beta_k)=p(z_k)$と

した。

【M ステップ】

パラメータ（$\boldsymbol{\alpha}(k), \boldsymbol{\beta}$）が与えられたときに観測変数と隠れ変数の組（$\boldsymbol{I}, \boldsymbol{Z}$）の結合対数尤度の期待値として定義される関数 Q を最大化する。ここで前段の E ステップで計算した **Z** の事後確率を用いる：

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= E\left[\ln\left\{p(\boldsymbol{I}, \boldsymbol{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta})\right\}\right] \\ &= E\left[\ln\left\{\prod_k \left(p(z_k)\prod_i p(I_{ik}|z_k, \alpha_i, \beta_k)\right)\right\}\right] \\ &= \sum_j E\left[\ln\left\{p(z_k)\right\}\right] + \sum_{ik} E\left[\ln\left\{p(I_{ik}|z_k, \alpha_i, \beta_k)\right\}\right] \end{aligned} \tag{3.90}$$

ただし、$\boldsymbol{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ があたえられたとき、I_{ik} は条件付き独立である。

ここで、前段の E-ステップで推定した α_{old}, β_{old} を用いて Z の期待値を計算する。勾配上昇法（gradient ascent を）用いて Q を局所的に最大化する α、β の値を見つける。

ここまではフレーム内における巡回的な最適化計算である。

3.20.5 トラッキングの性能改善

上記は複数の認識または計測結果をもとに予測した結果と新たに認識した結果との差分を評価するというプロセスを繰り返すことで認識器の性能をフレーム内で改善させようとするものであった。したがって、フレーム間の認識性能については依然として各認識器の個別の性能に頼らざるを得ない。

これに対し、近年、ラベリングそのものをトラッキングすなわちフレーム間認識のレベルで改善する試みが行われている。その手法は以下の３つに大別される：

（１） マルコフ確率場（MRF）を用いる手法（東大上条研など）

（２） モンテカルロフィルタ、パーティクルフィルタによる手法（INRIA 他）

（３） 上述の GLAD 手法を時間方向に拡張した手法[H2]（ハルビン工科大他）

ここでは複数の認識器を協調させるという点で前節の発展形に位置づけられるとの観点から、（３）に着目してその概要を示す。

いま、m 個の不完全な認識器が動画像の入力列 $\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_T$ を観測すると考える。ただし、

$$\boldsymbol{S}_t = \left\{\boldsymbol{x}_t^{\,1}, \boldsymbol{x}_t^{\,2}, \cdots, \boldsymbol{x}_t^{\,N_t}\right\} \tag{3.91}$$

はフレーム t で得られたサンプル集合であり、$\boldsymbol{x}_t^{\,k}$ は $\boldsymbol{S}_t$ における k 番目のブロック画像 $BLK(k)$ から構成される特徴ベクトルに相当する。$\boldsymbol{x}_t^{\,k}$ に対するクラスインデックスの候補集合を次式で記述する。

$$L_t^k = \left\{I_t^{i\hat{k}} \middle| \hat{k} = k\right\} \tag{3.92}$$

ただし、$I_t^{i\hat{k}} \in \{0,1\}$ は i 番目の認識器によって与えられる。弱い学習によるトラッカーの設定は以下のように行う。

以下の手順を *t=1,2,…*について繰り返す。

Step-1) サンプル集合 $S_t = \{x_t^1, x_t^2, \cdots, x_t^{N_t}\}$を受け取る

Step-2) この段階ではまだ S_{t-1}までしか学習していない不完全認識器群を用いて，各 x_t^k について $L_t^k = \{l_t^{i\hat{k}} | \hat{k} = k\}$を予測する

Step-3) S_tの正解インデックスと上記予測結果を比較し、各不完全認識器の認識精度(accuracy) を算出する。

Step-4) この認識精度と正解インデックスを S_tとともに学習する。

Step-5) ロバストなトラッキング結果であれば不完全認識器群を更新する。

学習データは *heuristic* に選ばれるが、学習完了後はあらたな学習データがなくても、入力に対して最も妥当な bounding box の位置ベクトルと各トラッカーのトラッキング精度を同時に推定する。

m 個の不完全認識器が時刻 t のフレームにおける現在の探索窓においてブロック画像集合から得られる $S_t = \{x_t^1, x_t^2, \cdots, x_t^{N_t}\}$を受け取るとする。ここで、 x_t^k は $BLK(k)$ から構成される特徴ベクトルである。 $P_t = \{p_t^1, p_t^2, \cdots, p_t^m\}$を時刻 t のフレームについて不完全トラッカー群 $T = \{T_1, T_2, \cdots, T_m\}$が出力したトラッキング結果（bounding box で表示）の集合とする。また、 $D(a,b)$を 2 つのブロック画像 a,b 間の中心位置の距離（画素）とする。そしてトラッカー T_iが出力した x_t^k のラベルを $l_t^{i\hat{k}} \in \{0,1\}$と記述する。各トラッカー T_iについてフレーム *t* の正例に属する上位 N_{BEST} 個の確信度（トラッカー T_iで推定される）を有するブロック画像の集合を Top_t^iで表す。 S_t中の各ブロック画像についてトレーニング集合（学習データ）に入れるかどうかの選別を以下の手順で行う：

$$
\begin{aligned}
&\text{For } k = 1,\ldots,N_t\{\\
&\quad \text{If } \left(p(x_t^k) \in P_t = \{p_t^1, p_t^2, \cdots, p_t^m\}\right)\\
&\quad \{\\
&\qquad \text{For } i = 1,\ldots,m\{\\
&\qquad\quad \text{If}\left(D\left(p(x_t^k), p_t^i\right) \le D_{th} \middle| x_t^k \in Top_t^i\right)\{l_t^{ik} = 1\}\\
&\qquad\quad \text{else } \{l_t^{ik} = 0\}\\
&\qquad\quad \text{endif}\\
&\qquad \}\\
&\quad \}\text{else}\{\text{neglect the image patch } x_t^k\}\\
&\}
\end{aligned}
$$

ここで， $p(x)$は特徴ベクトル x を構成するブロック画像の位置情報ベクトルを表す。文献[H2]では N_{BEST}=10、Dth=5 と設定されているが、採用した認識器や問題に応じて変わると考えられる。

さらに、ブロック画像 x_t^k の判別容易性 β_t^kを *m* 個の不完全認識器による多数決法で初期化すると次式のようになる：

$$\beta_t^k = \exp\left(\kappa_1 \times \left|\frac{2}{m}\sum_{i=1}^{m} l_t^{ik} - 1\right|\right) \tag{3.93}$$

ここで、 κ_1は正規化係数である。

m 個のトラッカーの精度 α_i^t，(i=1,…,*m*)について時間方向の更新は次式で行われる。

$$\alpha_i^t = (1 - w_i^t)\alpha_i^{t-1} + w_i^t M_i^t \tag{3.94}$$

ただし、w_i^t は学習率であり、また、M_i^t は正解との比較により成功したと判断されるトラッカーについて１、それ以外のモデルでは０となる。誤った更新を避けてロバスト性を最大化するために、学習率の設定は次式のように適応的に行う：

$$w_i^t = \begin{cases} \dfrac{\kappa_2}{1+e^{-\alpha_i}}, \text{if } M_i^t = 1 \\ \dfrac{\kappa_2}{1+e^{\alpha_i}}, \text{ if } M_i^t = 0 \end{cases} \tag{3.95}$$

ただし、κ_2 は正規化係数であり、文献[H2]では 0.1 に設定されている。トラッカーi が成功して精度 α_i が大きいことがわかれば、学習率 w_i^t を大きい値に設定して迅速に適応させる。成功しない場合は w_i^t を小さくする。

以上により学習段階が完了し、以後、各トラッカーにおいて予測が実行できる。

3.21 複数の認識エージェント

時系列ステレオの類推から、過去のデータに基づく映像予測は複数の視点による認識と同等に考えることができる。ここで、個々の認識器の動作を単なる画像センシングのみならず、他の認識器との協調、さらには情報交換によるマクロ概念の形成という動作としてとらえれば、個々の単体は認識器というよりも認識エージェントと呼ぶほうがふさわしい。複数の認識エージェントの協調には次のような効果がある。

１） 認識率を向上させる

２） 時空間に伴う変化を予測する

３） 新しい認識属性を獲得する（複数の 2 次元画像が立体構造を再構成するように）

もちろん、認識が完結する時間に制約を設けなければ、単体のエージェントが観測を繰り返すことで上記と同等の効果を得ることは可能である。しかし、リアルタイムに複数の場所に存在することができなければならないようなケースには対処できない。

認識エージェントの協調には様々な応用例がある。ロボット、プローブカー、Web 上の人工知能、人間社会における知の形成の説明などがそれにあたる。本論文ではこれらの説明は割愛し、MBV における動画像認識に焦点をあてる。

参考文献

A. 筆者文献

[1] M. Sasaki, “A Proposal for High-Efficiency Coding of the Definite Image Sequence”, *PCSJ88*.

[2] 笹木,“モデルベース顔領域抽出と領域別量子化による顔画質向上”,1997 年映像情報メディア学会年次大会 (ITE’97),p.381-382,1997 年 7 月.

[3] M.Sasaki, A.Toyoda, “3D Motion Estimation of a Moving Object using Model-Based Techniques”, ITS95 pp.1075-1081, November 1995.

[4] 笹木, 難波, 粉川, 小川, “車載情報通信におけるメディア適応化手法の提案 - 制約充足問題を用いた適応化手法の検討 - ”, 信学技報 ITS2004-22, Vol.104 No.325, pp.35-42, 2004 年 9 月

[5] N. Day, S. Sekiguchi and M. Sasaki, “21 Mobile Applications” in *Introduction to MPEG-7* edited by B. S. Manjunath et al. , pp. 353-361, WILEY, 2002.

[6] 笹木美樹男,難波秀彰： “ＩＴＳにおける知的情報支援とそれを支える視覚通信複合技術 - 知識共有型の環境認識をめざして - ”,2006 年度人工知能学会全国大会(JSAI 2006) 1E2-2, 2006 年 6 月.

[7] 笹木美樹男： “統計的推論に基づく景観認識の検討”,情報処理学会研究報告,2006 AVM-54-8 (2006).

[8] 笹木美樹男： “色情報と知識処理に基づく車載カメラ映像のインデキシング”,デンソーテクニカルレビュー Vol.12 No.1 2007, p.58-68.

[9] Mikio Sasaki, Kenji Muto, Kunihiko Sasaki: "Estimation of Driving Safety from On-Board Captured Scene by using Probe Image Database",*ITST2007*, June 2007.

[10] Mikio Sasaki, Hideaki Nanba: "Pedestrian Tracking by using MPEG-based Video Signal Processing", *IEICE Technical Report*, vol. 108, no. 344, IE2008-106, pp. 17-22, Dec. 2008.

[11] Mikio Sasaki: "MPEG-based Scene Class Recognition", *IEICE Technical Report*, vol.109, no.148, IE2009-55, July 2009.

[12] 笹木美樹男 :“MPEG 映像に適した景観認識手法の提案”,画像電子学会誌, vol.38, no.6, pp.890-899 (2009-11).

[13] Mikio Sasaki: "MPEG-based Scene Class Recognition", http://homepage3.nifty.com/MikioSasaki/scene/090724_IEICE_SceneClass.pdf

B. 画像符号化・映像メディア処理

[1] “Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 1: Systems”, ISO/IEC11172-1 First edition 1993-08-01.

[2] “Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 2: Video”, ISO/IEC11172-2 First edition.

[3] 石川, 相澤, 羽鳥, “簡易モデルのあてはめによる 3 次元仮想空間の構築手法”, 画像符号化・映像メディア処理シンポジウム(PCSJ・IMPS98) I-1.4, pp.9-10, Oct.26-28, 1998.

[4] ISO/IEC 15938-3 Information technology – Multi-media Content Description Interface Part 3: Visual.

[5] ISO/IEC 15938-5 Information technology -- Multimedia Content Description Interface Part 5: Multimedia Description Schemes.

[6] B.S.Manjunath *et al.*, “Introduction to MPEG-7”, WILEY, 2002.

[7] Youiti Horry, Ken-ich Anjyo and Kiyoshi Arai, “Tour into the picture”, Proc. SIGGRAPH’97, Los Angels, California, pp.225-232, August 3-8, 1997.

[8] 岩崎敏紀, 横山貴紀, 渡辺俊典, 古賀久志, “**M**PEG ビデオデータの動きベクトルを用いた圧縮領域における移動物体の検出と追跡”, 電子情報通信学会論文誌 D, Vol. J91-D, No. 6. (June 2008), pp. 1592-1603.

C. シーン理解

[1] Hoiem, A. A. Efros and M. Hebert, “Recovering Surface Layout from an Image”, *International Journal of Computer Vision* 75(1), 151-172, 2007.

[2] G. Pirirou, P. Bouthemy, and J.-F. Yao: “Recognition of Dynamic Video Contents With Global Probabilistic Models of Visual Motion”, *IEEE Trans. Image Process.*, vol. 15, no.11, pp.3417-3430, Nov. 2006.

[3] “DIGITAL VIDEO RETRIEVAL at NIST - TREC Video Retrieval Evaluation”, http://www-nlpir.nist.gov/projects/trecvid/

[4] J. Vogel, B. Schiele, “Semantic Modeling of Natural Scenes for Content-Based Image Retrieval”, *International Journal of Computer Vision* 72(2), 133-157, 2007.

[5] Anna Bosch, Andrew Zisserman, and Xavier Mun˜oz, “Scene Classification Using a Hybrid Generative/Discriminative Approach”, *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol.30, no.4, pp.712-727, Apr. 2008.

[6] Joakim Jitén Söderberg: “Multidimensional Hidden Markov Model Applied to Image and Video Analysis”, *A Thesis for the degree of DOCTOR OF PHILOSOPHY,* Institute Eurecom Sophia Antipolis, February 2007.

[7] Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[8] Ashutosh Saxena, Min Sun, and Andrew Y. Ng, "Make3D: Learning 3D Scene Structure from a Single Still Image", *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol.31, no.5, pp.824-840, May 2009.

[9] D. Hoiem, A. A. Efros and M. Hebert: “Recovering Surface Layout from an Image”, International Journal of Computer Vision 75(1), 151-172, 2007.

[10] “DIGITAL VIDEO RETRIEVAL at NIST - TREC Video Retrieval Evaluation”, http://www-nlpir.nist.gov/projects/trecvid/

[11] 浅田，白井，“マルチセンサ情報を動的に統合することによる道路シーンの解釈とモデリング”，情報処理学会論文誌，Vol.31, No.12, pp.1743-1754 （Dec.1990）.

[12] 杉山，目黒，金子，“少数の透視投影画像に基づいて生成された仮想空間における 3 次元移動表現”，信学技報 MVE 2001-148 (2002-3).

[13] Vittorio Ferrari, Andrew Zisserman, “Learning Visual Attributes”, http://eprints.pascal-network.org/archive/00003146/01/ferrari07.pdf

[14] Bangpeng Yao, Gary Bradski, Li Fei-Fei, “A Codebook-Free and Annotation-Free Approach for Fine-Grained Image Categorization, http://ai.stanford.edu/~bangpeng/document/cvpr12_0654.pdf

D. 因子分解法

[1] 金出武雄，コンラッド・ポールマン，森田俊彦，“因子分解法による物体形状とカメラ運動の復元”，電子情報通信学会論文誌 D-II，J74-D-II-8, pp. 1497-1505, Aug.1993.

[2] J. P. Costeira, “A Multibody Factorization for Independently Moving Objects”, *International Journal of Computer Vision* 29(3), 159-179(1998).

E. 歩行者検出・トラッキング

[1] Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[2] P. Viola, M. J. Jones, “Robust Real-time Object Detection”, *Cambridge Research Laboratory Technical Report Series CRL2001/01, Feb. 2001*.

[3] G. Antonini, “A discrete choice modeling framework for pedestrian walking behavior with application to human tracking in video sequences”,*Thèse EPFL*, no 3382 (2005).

[4] http://www.gavrila.net/Research/Pedestrian_Detection/Pedestrian_Projects/SAVE-U_Final_Report.pdf

[5] Xiang, Tao; Gong, Shaogang; “Beyond Tracking: Modeling Activity and Understanding Behaviour.“, International Journal of Computer Vision, Vol. 67, No. 1, April 2006 , pp. 21-51(31).

[6] J.-M. Odobez, D. Gatica-Perez, and S. O. Ba, “Embedding Motion in Model-Based Stochastic Tracking”, *IEEE Trans. Image Processing*, vol.15, no.11, pp.3514-3530, Nov. 2006.

[7] E. Arnaud, E. Mémin, “Partial Linear Gaussian Models for Tracking in Image Sequences Using Sequential Monte Carlo Methods”, *International Journal of Computer Vision* 74(1), 75-102, 2007.

[8] Bo Wu, Ram Nevatia, “Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors”, *International Journal of Computer Vision* Volume 75, Issue 2(November 2007), 247 – 266.

[9] M. Lyell, R. Flo, M. Mejia-Tellez,”Simulation of Pedestrian Agent Crowds, with Crisis”, http://necsi.org/events/iccs6/viewpaper.php?id=367.

[10] D. M. Gavrila, S. Munder, “Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle”, *International Journal of Computer Vision* 73(1), 41-59,2007.

[11] J. Tilp, et al., “Pedestrian Protection based on Combined Sensor Systems”, Paper Number 05-0156, http://www-nrd.nhtsa.dot.gov/pdf/nrd-01/esv/esv19/05-0156-O.pdf

[12] Ahmed Elgammal, Chan-Su Lee, “Tracking People on a Torus”, *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol.31, no.3, pp.520-538, Mar. 2009.

F. 一般物体認識

[1] 神谷保徳，矢野良和，大熊繁，高橋友和，井手一郎，村瀬洋：”一般物体認識のためのタイプの異なる局所特徴の統合利用”，信学論 D, Vol. J92-D, No.5, pp628-638 (2009)

[2] 三浦純，森田英夫，ヒルドミヒャエル，白井良明：”SVM による物体と位置の視覚学習に基づく屋外移動ロボットの位置推定．日本ロボット学会誌，25(5), pp. 792−798, 2007.

[3] Wan-Lei Zhao, Chong-Wah Ngo, “lip-vireo Manual Version 1.01”, http://www.cs.cityu.edu.hk/~wzhao2/lip-vireo.htm

G. 予測・推定・多変量解析・時系列分析

[1] J.-S. R. Jang: “ANFIS: Adaptive-Network-Based Fuzzy Inference System”, *IEEE Trans. Syst., Man, Cybern.*, vol. 23, no.3, pp.665-684, May/June 1993.

[2] Y. Takane, Michael A. Hunter, “Constrained Principal Component Analysis: A Comprehensive Theory”, AAECC 12, 391-419 (2001).

[3] 滝口哲也，有木康雄，“ Kernel PCA を用いた残響下におけるロバスト特徴量抽出の検討，” 情報処理学会論文誌，Vol.47，No.6，pp. 1767-1773，2006．http://www.me.cs.scitec.kobe-u.ac.jp/~takigu/pdf/2005/kpca.pdf

H. GLAD 法

[1] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan, “Whose Vote Should Count More:Optimal Integration of Labels from Labelers of Unknown Expertise”, http://mplab.ucsd.edu/~jake/OptimalLabeling.pdf

[2] B. Zhong, et al., “Visual Tracking via Weakly Supervised Learning from Multiple Imperfect Oracles”, *CVPR2010*.

I. ITS 関連

[1] K. Muto, K. Matsugatani, M. Egawa, and H. Nanba, “Omni-Directional Picture Database Using Probe Cars”, *12th World Congress on ITS*, 6-10 Nov. 2005.

[2] Thaned SATIENNAM, “A Study on Pedestrian Accidents and Investigation of Pedestrian’s Unsafe Conditions in Khon Kaen Municipality, Thailand”, *Journal of the Eastern Asia Society for Transportation Studies*, Vol.5, October, 2003.

[3] 関本義秀，菊地英一，佐藤圭一，秋山祐樹 ： パーソントリップデータを活用した人の流れの時空間的な詳細化，第 28 回交通工学研究発表会論文集，pp.197-200, 2008.

[4] http://www.dft.gov.uk/pgr/roads/vehicles/intelligentspeedadaptation/

[5] 熊谷，伏木，横田，君田，“特徴空間射影によるプローブカーデータのリアルタイム補完”，情報処理学会論文誌，Vol.47, No.7, pp.2133-2140 (2006).

[6] T. Fushiki, et al, “Traffic Condition Prediction by Use of Floating Cars”, *IPSJ Journal* Vol. 43 No. 12, pp. 3801-3808, Dec. 2002.

[7] M. Kumagai, et al, “Efficient Forecast Process for Nationwide Traffic Information Services”, *Technical Report Of IEICE* ITS2004-20(2004-09)

[8] T. Fushiki, et al, “Traffic Condition Prediction by Use of Floating Cars”, *IPSJ Journal* Vol. 43 No. 12, pp. 3801-3808, Dec. 2002.

J. その他

[1] 北，津田，獅々堀，“情報検索アルゴリズム”，共立出版，2002.