論　文

属性に基づく学習型人物検索

井尻　善久[†,††a)]　勞　世竑[†]　村瀬　洋[††]

Learning Based Human Image Search Using Facial Attributes

Yoshihisa IJIRI[†,††a)], Shihong LAO[†], and Hiroshi MURASE[††]

あらまし　人物属性による人物検索の新たなフレームワークを提案する．本論文における人物属性とは，性別・表情等に加え眼鏡等装着物の有無等を含む顔周辺に現れる属性である．人物属性は多様であり，更にそれらの組合せの属性が存在する事を考慮すると，その種類は膨大であり，あらかじめ考えられる全属性の識別器を構築しておく手法には限界がある．これに対し本論文では，検索時に検索に用いたい属性を含む学習サンプルを数枚用意し，属性識別器を構築し，検索を行う学習型検索手法を新たに提案する．これにより，学習サンプルさえ用意すれば様々な属性を用いて検索ができるようになる．一方，識別器の学習は検索時に行われるので，高速な学習手法が求められる．このためあらかじめ属性を構成する局所的パターン集合を構築しておき，検索時には，これらパターンと学習サンプルを用いた特徴の抽出と，属性の学習のみを行う．これにより，検索時に高速に新たな属性を学習できるようになった．提案手法の有効性は実験により実証する．

キーワード　人物，顔，属性，検索

1. まえがき

公共の場での安心感や安全性の向上のために，多くの監視カメラが公共の施設に設置されるようになってきており，大量の映像を処理する必要が生じている．また，一般人の身の周りでも，ディジタルカメラ等の普及から以前とは比較にならないほど映像や画像が氾濫するようになってきた．これに伴い，大量の画像の中から特定の人物を見つけるのが困難になってきており，何らかの支援システムが必要となってきている．

こうした必要性は比較的身近なところで実感できる．例えば捜査の現場においてはある監視カメラに映った犯人の映像をもとに同様の特徴をもった人物を大量の記録から検索するのに多大の工数がかかっている．また，撮りためた動画や静止画から，自分の子供や友人等，特定人物を素早く見つけたいと思うことはしばしばである．これに対し顔認識が提案されてきたが，顔が鮮明に見えない状況等では十分な精度を発揮しにくい（図 1）．また，特定したい人物の顔画像が入手できない場合には使えない．こうした条件下で特定の人物を探し出すには，性別・表情や眼鏡・サングラス・マスク等装着物有無の「人物属性」を手掛りとした方が有効である場合も多い．したがって本論文においては人物属性を用いて人物検索することを考える．

人物属性に基づく人物検索において，検索に用いられる可能性のある人物属性は，非常に多様である．例えばあるときはサングラスをかけた人物という形で絞り込みたいかもしれないし，別のときには，金髪の人物や，灰色のニット帽をかぶった人物を見つけたいか

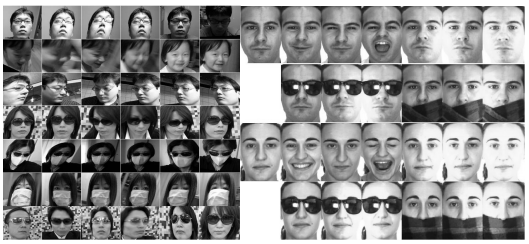

図 1　対象とする多様な属性の例：(左) OMRON DB, (右) AR DB
Fig. 1　Examples of target attributes: (left) OMRON DB, (right) AR DB.

† オムロン株式会社技術本部コアテクノロジーセンタ，木津川市
Core Technology Center, Corporate R&D, OMRON Corp., 9–1 Kizugawadai, Kizugawa-shi, 619–0283 Japan
†† 名古屋大学大学院情報科学研究科，名古屋市
Graduate School of Information Science, Nagoya Univ., Furo-cho, Chikusa-ku, Nagoya-shi, 464–8603 Japan
a) E-mail: yoshihisa_ijiri@omron.co.jp

もしれない．このように用いられ得る属性の種類は多様であり，更にそれら属性の組合せが発生する可能性があることを考えると，対応しなければならない属性の数は非常に多くなる．したがって，仮にあらかじめ構築した属性識別器を用いる手法をとった場合，使われる可能性のある属性すべてに対応する膨大な数の識別器を用意しなければならなくなり，現実的とはいえない．そこで，本論文では，検索キーとして用いたい属性を有した人物のサンプル画像を利用する手法を考える．この場合，サンプル画像を用意する必要があるが，それさえ用意すればどんな属性でも検索が可能となる．実用的には多くの条件で簡単にサンプル画像を用意することができる．例えば，前述の捜査の現場においては，監視カメラに映った数枚の犯人映像を利用できる．また，家族アルバムから，笑顔の写真等，特定の属性の顔を数枚〜数十枚用意することはさほど難しくない．したがって，本論文で論じる，サンプル画像を用いた検索は現実的である．

従来のサンプル画像を用いた検索においては，サンプル画像と検索対象画像の距離指標に基づいて一定の距離以内にある画像を検索結果として返す方法が一般的である．本論文では，より高い精度で検索を行うために，検索要求時に短時間で属性を学習する「学習型検索」を提案する．この際に問題となるのは，学習サンプルとして必要な画像枚数と，学習に要する時間である．数百枚もの学習サンプルを用意して，検索を行うのは容易ではない．また，学習に時間がかかれば，検索要求時に待たされることになるので，使い勝手が損なわれることになる．したがって学習時間は，人が待てる時間内（数秒程度）であることが重要となる．以上の要件をまとめると，学習型人物検索においては，少数の学習サンプルにより短時間で属性の学習が行えることが要件となる．

一方，属性認識に関しては多くの顔に基づく認識手法が既に提案されている．Moghaddam [1]，Hosoi [2]，Hayashi [3]，Zhuang ら [4] は性別，年齢，人種等を個別に推定している．これに対し Lyon ら [5] は性別，人種，表情の複数属性に同時対応し，Wilhelm ら [6] は性別，年齢，表情の推定に加え顔認識を同時に実現している．これら初期の手法は，あらかじめ構築した特定の属性に特化した属性識別器を用いる方法であり，学習された属性において高い精度を発揮するが，本研究が目的とするような多様な属性検索に応用することはできない．

比較的最近では多様な人物属性に対応したフレームワークとして Kumar ら [7] が次のような手順からなる Face Tracer を提案している．(1) あらかじめ決めた 10 個の局所領域で合計 450 種類の画像記述（色空間，正規化方法等の組合せ）を行う，(2) それぞれに対応する 450 個の Support Vector Machine（SVM）を訓練する，(3) Adaboost により精度の高い SVM のみを選択する，(4) 選択された SVM の出力を更に上位の SVM で統合する（図 2 左）．この手法の特長は，特定の属性に特化した特徴量を用いる代わりに 450 種類の冗長な画像記述を行うことにより様々な属性に対応できることである．この手法も，あらかじめ構築した特定の属性識別器を用いる手法であるが，様々な属性を学習できるフレームワークであることが，初期の方法より優れている点である．また，あらかじめ大量の学習データを用いて識別器を学習することができるので，学習した属性に対しては高い精度を発揮する．しかし，この手法を学習型検索に応用する場合には，検索要求時に 450 種類の画像記述・識別器学習及び識別器選択・統合を，一から行わなければならないので，時間がかかる（後に示す実験結果によれば数十分〜数時間程度）．結果的に，検索を行う際に，長時間待ち時間が発生することになり効率的ではない．したがってこの手法も，本研究が目的とする学習型人物検索という観点では適しているとはいえない．

Kumar らの手法では，どの特徴が属性認識に適しているのかが事前に分からない前提であるので，冗長に特徴を抽出し，それらから適したものを選択する方法となっている．逆にどのような特徴が属性認識に役

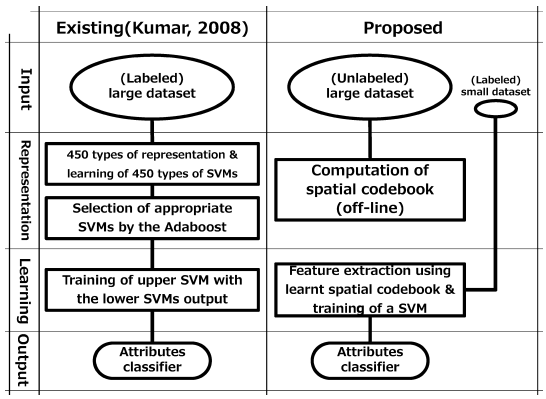

図 2　提案手法と従来手法における学習フローの対比
Fig. 2　Comparison of learning flow between proposed and current state-of-art method.

立つのかがあらかじめ分かっていれば，少数の学習サンプルからでも効率的にその特徴を取り出すことができ，また短時間で新たな属性を学習できるはずであるというのが，提案手法の考え方の基本である．このため，提案手法では設計段階で，（大量の）ラベル無データを用いて，生じ得る（ラベル無データに含まれる）人物属性を構成する局所的パターンを集約した spatial codeword の集合（spatial codebook）を構築する．局所的パターンを用いるのは，人物位置が特定されていることを前提とすれば，人物属性は注目領域内で比較的局所的に現れることが多いためである．一方，検索要求時には，検索したい属性の少数の学習サンプルを positive，及びそれと判別したい画像を negative として与え，それらサンプルと各 spatial codeword（生じ得る局所パターン）との相違度（距離）を特徴量として属性識別器の学習を行う（学習型検索）．実験結果により示すが，この学習は，Kumar らの手法が大量のデータを必要とするのに対して，比較的少ない学習サンプル（数十枚程度）により行うことができる．これにより，Kumar らの手法で数時間要した学習が，数秒で可能となり学習型人物検索に応用できるようになった．このように，提案手法は，学習ステップを，属性認識に役立つと思われる特徴量を集約し典型的な局所パターンを事前に抽出しておく段階と，それら局所パターンとの類似性をもとに属性学習を行う段階の，2 段階に分割し，後者のみを属性検索時に行うので，前者と後者を同時に行わなければならない従来手法に比べ学習が効率的であり（図 2 右），これにより検索時の学習が可能となっている．Kumar らによる手法と，提案手法の対比は表 1 に示した．

なお，提案手法の具体的な使い方は次のような流れになる．例えば，眼鏡の人物を検索したい場合には，眼鏡をかけた人物とかけていない人物のサンプル画像を数十枚集める．次いで，それらの画像をシステムに入力して，検索ボタンを押す．内部的に学習型検索が動作し，数秒（対象データの多さに依存するが 1 万枚程度の場合）で高精度な検索結果がユーザに提示される．

以上をまとめると，提案手法の新規性は，従来のように特定の属性のみをあらかじめ学習しておく方法と異なり，検索時に少数の学習サンプルを用いて短時間で行う学習型人物検索により，任意の属性による高精度な検索を可能としたことである．提案手法の有効性は実験により実証した．本論文の残りの部分は，次のように構成される．まず **2.** において提案手法について詳述し，**3.** で提案手法の有効性を示すための実験結果について述べる．次いで，**4.** において簡単な考察をした後に，最後に **5.** で結論を述べる．

表 1　提案手法と従来手法の対比
Table 1　Comparison between proposed and existing methods.

項目	従来手法	提案手法
学習データ	大量のラベル有データ	大量のラベル無データ 少数のラベル有データ
学習時間	長時間 （数十分〜数時間）	短時間 （数秒）
学習できる属性	特定属性（性別等） Kumar [7] は拡張可	ユーザ定義で拡張可能
精度	学習データが多いとき高精度	学習データが少なくても比較的高精度
利点	使う属性が決まっているときに高精度	用いたい属性を随時追加変更できる

2. 提案するフレームワーク

以降の記述では，人物領域が顔検出若しくは人物検出等によりある程度正確に抽出されていることを仮定する．設計段階に，様々な属性を有した人物を含む大量のラベル無データ $X = \{x_n; n = 1, \ldots, N_x\}$ を用い，属性を構成する顔内部の局所的基本パターンを表現する spatial codeword $\{s_t; t = 1, \ldots, T\}$ からなる集合 spatial codebook S を構築する．更に，オフラインで検索対象画像 $Z = \{z_n; n = 1, \ldots, N_z\}$ に対して，各 spatial codeword と各学習サンプルとの距離 $\{f_t(z_n); t = 1, \ldots, T\}$ を識別のための特徴(注1)として抽出しておく．検索要求時には，検索したい属性を含む学習サンプルを用意し，検索対象画像に対すると同じように各 spatial codeword と各学習サンプルとの距離 $\{f_t; t = 1, \ldots, T\}$ を識別のための特徴量として抽出し，SVM で学習することにより，識別器 H を学習する．次いで，学習した識別器 H を，各検索対象画像から抽出された識別のための特徴 $f_t(z_n)$ に対して適用し，各検索対象画像に対する検索結果のスコア $H(z_n)$ を得る．このスコアにしきい値を適用し，しきい値以上であれば検索結果として提示する．各ステップについて以降の各節で詳述する．

(注1)：本論文では，spatial codebook 構築の際，特別な特徴量を用いずグレースケール入力画像を直接利用した．しかしながら，色情報等を利用した特徴量等を用いることもできる．この際の特徴抽出と，識別器に直接入力される特徴量を区別するため，ここでは「識別のための特徴」という用語を用いる．

2.1 Spatial codebook の構築方法

多様な人物属性を表現するために，それらの属性を構成する局所的基本パターン集合を大量のラベル無データ X から抽出する．局所的基本パターンは，人物領域を局所領域に分割し（領域間の重なりがあってもよい），その各局所領域において，ラベル無データ X に K-means クラスタリングを適用することにより抽出する [8], [9]．ここで局所領域の数を L，各局所領域におけるクラスタ数を K とすると，全体で KL 個のクラスタ $\{C_{lk}; k = 1, \ldots, K; l = 1, \ldots, L\}$ を得る．KL 個すべてのクラスタを用いることは冗長である．例えば，左目の領域でサングラスのクラスタを抽出しておき，右目の領域で同じくサングラスのクラスタを抽出したとすれば，これらのクラスタのメンバとなるデータはほぼ同じになる可能性が高い．このような場合，どちらかを除去してもサングラスのクラスタが出現したことを表現できる．したがって，同じようなクラスタリング結果になるようなクラスタを除去し，選択された各クラスタが表現できるデータが互いに補完的になっているようにすることが望ましい．このため KL 個のクラスタからそれを用いてクラスタリングされるメンバができるだけ直交となるようにクラスタを選択することを考える．また一方で，ほとんどのデータをメンバとして含むようなクラスタは，属性を見分けることを考えたとき効率が悪い．またメンバ数が極端に多い若しくは少ないクラスタは一つの属性というよりは全体の傾向や特殊なノイズを表現していることも多い [10]．逆にそうしたクラスタを除去した後に残るクラスタは何らかの属性の構成要素である可能性が高い．

こうした条件を満たす選択を実現するために，各クラスタ C_{lk} に対して，そのクラスタへの X の各データのメンバシップを表す $\phi_{lk}(x_n)$ を次のように定義する．

$$\phi_{lk}(x_n) = \begin{cases} 1 & x_n \in C_{lk} \text{ のとき} \\ 0 & x_n \notin C_{lk} \text{ のとき}, \end{cases} \tag{1}$$

ここで，$l = 1, \ldots, L$ は局所的な位置，$k = 1, \ldots, K$ は各局所領域における各クラスタ，$n = 1, \ldots, N$ でありラベル無学習データ x のインデックスである．これを用い，個別の属性の構成要素である可能性が高いクラスタを選択するために，ϕ_{lk} のエントロピー $-\sum P(\phi_{lk}) \log(P(\phi_{lk}))$ が最大（maximum entropy）であるものを選択する．また，できる限り個別の属性を抽出するために，既に選択された $\phi_{l\tilde{k}}$ に直交する ϕ_{lk} をもつ C_{lk} を順番に選択していく．実際には選択が進むにつれ，完全に直交する ϕ_{lk} を求めることは困難になるので，内積が最小である C_{lk} を選択することにより最も直交性が高いものを選択する（maximum orthogonality）．これらの二つの基準を組み合わせたものを MEMO 基準（Maximum Entropy Maximum Orthogonality）と呼ぶことにする．MEMO 基準を使うことにより，独立な属性変動を表すクラスタを抽出することができる．Algorithm 1 に構築法をまとめる．MEMO 基準の有効性については後の実験で示す．この基準により選択された spatial codebook の一例を図 3 に示した．各画像内太枠が入力領域全体に対する spatial codeword の位置を表している．これを見ると，顔の比較的小さな局所領域のクラスタが codeword になっており，ある程度個別の属性の構成要素となっている様子が分かる．

図 3 Spatial codebook の一例：各ブロックが顔領域を表している．その中の更に小さな局所領域は選択された spatial codeword の位置を表している．

Fig. 3 An example of the spatial codebook: Each block stands for face region, where a smaller region in each block stands for a region to extract each spatial codeword.

2.2 Spatial Codebook を用いた識別のための特徴量抽出

本節では，前節で得られた spatial codebook による「識別のための特徴量」の抽出手法を説明する．従来の物体認識等においては，keypoint 検出器等により得られた局所領域パターンと最も近い codeword を一つ選択しそのクラスインデックスを特徴量とするのが一般的である．結果的に，例えば二つの codeword の

Algorithm 1 The spatial codebook based on the MEMO criteria

Require: 様々な属性を含む大量ラベル無データ X; spatial codeword の数 T;
Ensure: MEMO 基準を満たす spatial codebook S

1: 初期化：$S = \{\emptyset\}$
2: ラベル無データ X から L 個の局所領域を切り出し
3: **for** 局所領域 $l = 1, \ldots, L$ **do**
4: 　各局所領域 l において K-means を実行し K 個のクラスタ中心 $C_{lk}(k = 1, \ldots, K)$ を求める.
5: **end for**
6: ラベル無データ $x_n \in X$ に対して，式 (1) のメンバシップ $\phi_{lk}(x_n)$ を計算
7: 得られた LK 個のクラスタ中心 C_{lk} から $\phi_{lk}(X)$ のエントロピーが最大になるような $C_{\hat{l}_1\hat{k}_1}$ を選択し S に追加.

すなわち $(\hat{l}_1, \hat{k}_1) = \arg\max_{(l,k)} \left\{ -\sum_{x_n \in X} \mathrm{P}(\phi_{lk}) \log(\mathrm{P}(\phi_{lk})) \right\}$, $s_1 = C_{\hat{l}_1\hat{k}_1}$, $S = S \cup s_1$

8: **for** $t = 2, \ldots, T$ **do**
9: 　既に選択された S と直交性が最大でありエントロピーができる限り大きくなるような新しい C_{lk} を選択し S に追加.

すなわち $(\hat{l}_t, \hat{k}_t) = \arg\max_{(l,k)\setminus(\hat{l},\hat{k})} \left\{ \frac{1}{||C_{lk}^T S||} + \alpha \left(-\sum_{x_n \in X} \mathrm{P}(\phi_{lk}) \log(\mathrm{P}(\phi_{lk})) \right) \right\}$, $s_t = C_{\hat{l}_t\hat{k}_t}$, $S = S \cup s_t$. ただし

$(\hat{l}, \hat{k}) = ((\hat{l}_1, \hat{k}_1), (\hat{l}_2, \hat{k}_2), \ldots, (\hat{l}_{t-1}, \hat{k}_{t-1}))$, α は適当なバランスパラメータ.
10: **end for**

ちょうど中間に位置するような局所領域パターンに関しては，どちらかに強制的に割り付けられることになり，両方に類似しているという情報を表現できなくなる hard assignment 問題があった [11]〜[14]．従来の方法でこうした符号化誤差を解決するには，前述の中間位置にあるような局所領域パターンも含むように，より多くの codeword を用意する必要があり，結果として数千の codeword からなる冗長な codebook となる傾向があった．

これに対し，各 codeword と入力画像との距離に基づく特徴抽出により，画像検索等において，従来のベクトル量子化 [8], [9] に比べ高い精度が得られることが報告されている [11] ので，提案フレームワークにおいても各 codeword との距離に基づいて，識別のための特徴量抽出を行う．したがって，画像 ξ に対する識別のための特徴量 f は，次の式により得られる．

$$f_t = \mathrm{dist}(s_t, \xi(\hat{l}_t)). \tag{2}$$

ただし，$t = 1, \ldots, T$，$f \in \mathcal{R}^T$，$\xi(\hat{l}_t)$ は ξ の $\hat{l}_t$ の局所領域を表す．図 4 は，図内左側の各入力画像 ξ に対してどのように識別のための特徴量が抽出されるかを示している．横軸は特徴量のインデックス t を表しており，縦軸は得られた識別のための特徴量 f_t を表している．各横軸の下に示しているのは，距離の短かった spatial codeword である．結果を見ると入力した画像に類似の spatial codeword であることが分かる．検索を高速に行うためにはオフラインでできる処理をすべて行っておくことが必要である．そこで，検索対象画像 $Z = \{z_n; n = 1, \ldots, N_z\}$ に対し，識別のための特徴量 $f(Z)$ をオフラインで抽出しておく．

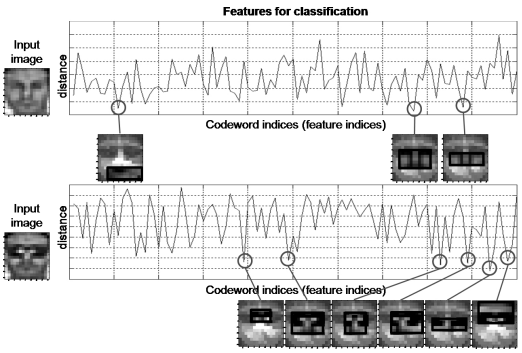

図 4　識別のための特徴量の一例
Fig. 4　An example of features for classifier inputs.

2.3　検索要求時における属性の学習

検索要求時には検索対象となる属性を有する数枚の学習サンプルと判別したい対象の学習サンプルから，その属性に特化した識別器を学習する．式 (2) により識別のための特徴量を抽出し，検索したい属性とそれ以外の属性を識別する識別器 H を構築する．識別器の構築方法としては様々な手法が考えられるが，本論文では最大マージン法に基づく SVM を利用した．

Face Tracer のように，ピクセルレベル若しくはそれと同程度の数百から数千次元からなる識別用の特徴量を用いた場合には，過適合を防ぐために，特徴量次

元数に見合った数の学習サンプルを用意しなければならない．しかし提案手法においては，数 10〜100 個に集約した spatial codeword を用いて特徴抽出を行うので，識別のための特徴量も数 10〜100 次元程度である．したがって学習サンプルは次元数と同程度の枚数（100 枚程度）で安定した性能を得ることができた．これにより，比較的少数の学習サンプルにより数秒間で新しい属性の学習が可能となった（学習時間は **3. 3** で詳述）．

2. 4 属性に基づく人物検索

検索時には，**2. 2** に述べた，検索対象の識別用の特徴 $f(Z)$ に対して，**2. 3** で学習した識別器 H を適用して，検索結果の集合 Z^* を次のように求める．

$$Z^* = \{Z; H(Z) > A\}, \tag{3}$$

ただし，A は適当なしきい値であり，A を大きくすると一般に再現率が低下し，小さくすると（同時に適合率が低下することも多いが）再現率が向上する．また，これを属性認識に応用する際には，判別対象である各属性 ω_i に対して「1 vs. all」型の学習を行うことにより多クラス識別器を構築し，次のように推定した．

$$\hat{\omega} = \arg\max_i \{H_i(Z; \omega_i)\}, \tag{4}$$

3. 実　　験

本章においては，提案手法の有効性を示すために，五つの実験を行う．まず最初に，(a) 提案した学習型人物検索により，様々な検索が高い精度で行えることを示す．次いで，「少数の学習サンプルで属性を学習できる」ことを示すために，(b) 学習サンプル数が少ないときの精度を既存手法と比較する．また，「属性を即座に学習できる」ことを示すために，(c) 学習に要する時間を既存手法と比較する．更に，Codebook の生成方法において，(e) 局所領域に注目することの有効性と，(f) MEMO 基準の有効性を検証する．検索の評価は，再現率・適合率・F 値等と指標が多くなるので，他手法と比較した基本的な特性を調べることが目的である (b)〜(f) の実験においては，性能変化が分かりやすい属性認識に応用して実験を行う．

3. 1 様々な属性による検索

提案手法により様々な属性の人物検索を行えることを示すために，AR face dataset（以降 AR DB）[15], [16] を用い評価を行った．AR DB は Martinez らにより撮影された顔認識用 DB であるが，通常の顔のほかサングラスやスカーフを装着した顔，また極端な表情として「叫び（Scream）」等が含まれているので，属性認識等の評価として用いた．この DB に含まれる顔の顔向きは正面のみであり，安定した照明条件下で撮影されている．また，各人 26 枚男女それぞれ 50 人の合計 2600 枚が含まれている．そこで，各画像を左右反転したものと併せ 5200 枚に拡充した後，無作為に 2600 枚を抽出し評価データとした．残りの 2600 枚は spatial codebook の構築に用いた．画像の一例は図 1 右に示している．一方，検索要求時の学習に用いた学習サンプルの枚数はいずれも 100 枚である（頻繁に検索する属性を変更するような状況で毎回 100 枚を集めるのは現実的ではないが，ある固定的属性を頻繁に用いる場合には一度集め保存しておけばよいので現実的な場合もある．更に 100 枚の学習サンプルは動画においては数秒で集めることができる）．また，この DB では Ground-Truth を用い 64×75 [pix]（両目間のサイズは 28 [pix] 程度）となるように正規化した．この DB に対して検索の実験を 10 回のクロスバリデーションにより行った結果を，表 2 に示した．なお，表中における異常照明条件は AR DB に含まれる標準的な照明条件と異なる照明条件である．示した結果は F 値が最も高くなるようにしきい値 A を調整したときの結果である．また，spatial codeword の数は $T = 100$ としている．性別による検索においては，スカーフやサングラスを装着した条件もテスト画像に含めているので精度が低くなっているが，おおむね 90%程度の F 値が得られており，高い検索精度が得られていることが分かる．

表 2　様々な属性検索の精度
Table 2　Search accuracies against various attributes.

検索条件	再現率	適合率	F-値
通常	88.5	80.1	84.1
笑顔	75.5	85.3	80.1
叫び	86.5	97.2	91.5
異常照明条件	94.6	98.4	96.4
サングラス	88.5	93.7	91.0
スカーフ	88.5	92.2	90.3
男性	87.2	88.2	87.7
女性	88.2	87.1	87.6

3. 2 学習サンプル数が少ないときの精度

提案手法が既存手法に比べ特に学習サンプル数が少ないときに有効であることを示すために，学習サンプルの枚数を変更しながら認識精度の比較を行った．対

象データとして，AR DB と OMRON 社内において収集された OMRON DB を利用した．AR DB における評価では (1) 装着物なし，(2) サングラス，(3) スカーフを装着した顔，(4)「叫び」の，4 クラス識別を実験した．一方，OMRON DB は，装着物なしの人物，サングラスやマスクを装着した多様な人物が含まれるデータベースである．それぞれの人物は多様な照明条件・表情・顔向きの条件で撮影されている．図 1 左はこの DB に含まれる画像の例を示している．この DB には顔位置等の人物位置を特定するための Ground-Truth が存在しないので OKAO Vision ライブラリ [17] の顔検出 [18]（サングラスやマスクを装着した顔でも顔検出可能）及び顔器官検出 [19] を用いて自動的に人物位置を特定した(注2)．したがって，この DB の実験結果は様々な顔の変動が生じたときの実用的な精度と考えることができる．この DB における評価データは 9068 枚であり，spatial codebook の構築には，評価用と異なる 2032 枚を用いた．本 DB におけるテストでは (1) 装着物なし，(2) サングラス，(3) マスク，(4) サングラスとマスクの両方を装着した人物の 4 クラス識別を行った．比較手法の Face Tracer の実装においては，提案されているグレースケール，カラー（RGB，HSV），エッジ方向・強度の 5 種類の画像記述のうち，提案手法とできる限り同条件で比較するためグレースケールのみを用いた．この結果，450 種類の画像記述から 90 種類となった．実験は各学習サンプル数で，10 回のクロスバリデーションにより行った．また，この実験においても spatial codeword の数は $T = 100$ とした．結果は図 5 に示したとおりである．これを見ると，特に学習サンプル数が少ないときに，提案手法が Face Tracer に比べ精度が高いことが分かる．また，学習サンプル数が少ないときにおける手法として，最も直接的なアプローチと考えられる正規化相関（表中 NCC）との比較も行ったが，この比較においても同様の結果となった．特に OMRON DB のように実環境に近い環境においては NCC より提案手法が高い精度となった．したがって，他手法と比較して少ない学習サンプル数で属性の学習ができることが実証された．

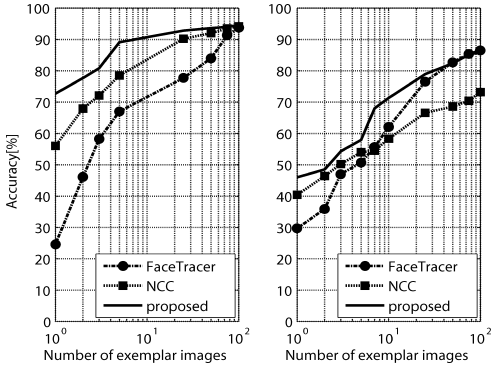

図 5　学習サンプル数と精度の関係；(左) AR DB；(右) OMRON DB

Fig. 5　Accuracies against training samples; (left) AR DB; (right) OMRON DB.

3.3　属性検索時の学習に要する時間

本節では，属性検索時の学習に要する時間を評価する．比較には AR DB を用いた．属性検索時に必要な処理は，提案手法では学習サンプルに対する識別用特徴量抽出と SVM の訓練である．一方，Face Tracer では，特徴量抽出及び各特徴量に対する SVM の学習，adaboost による SVM 選択及びその後段の SVM 学習である．実験においては，各属性に対し 100 枚の学習サンプルを用いた．また，FaceTracer の実装では前節の評価と同様に 90 種類のグレースケールの特徴量のみを用いた．また 90 個の SVM すべてについてパラメータ最適化するのは困難であるので，あらかじめ求めておいた経験的に良いパラメータを用いた(注3)．この条件で，属性検索時に必要な処理に要した時間を計測した結果は，表 3 に示したとおりである．従来手法においては一つの検索ごとに 30 分近く要しているのに対し，提案手法は 3 秒程度で学習が終了しており，検索時に属性を学習するに現実的な速度になっていることが分かる．従来手法の学習時間が長いのは，最終的に選択されない特徴量も含めて 90 種類の冗長な特徴量抽出を行い，そのすべてに対して SVM の学習を行っていることが，主な原因である．これに対して，提案手法はこの部分をオフラインで属性を構成す

表 3　検索時の学習に要する時間

Table 3　Run time for learning an attribute.

手法	学習時間 [s]
Baseline: Face Tracer	1515 程度
提案手法	2.875

(注2)：OKAO Vision ライブラリを用いると大きな顔向きの顔も顔検出されるが透視投影変換により顔と口の位置が一定になるようにした．目口検出が失敗した場合も評価に含めた（図 1 において顔向きが大きいときにひずんで見えるのはこのためである）．

(注3)：その他の実験条件；CPU：Core 2 Duo 3.00 GHz，メモリ：2 GByte，実装言語：MATLAB，SVM の実装には SVM^{light} [20] を使用．

表 4　局所抽出/全体抽出時の認識精度比較
Table 4　Comparison between local and holistic feature extraction.

手法	OMRON DB	AR DB
NCC	73.2±1.9	94.2±0.6
PCA + SVM	80.3±1.8	92.3±1.2
Face Tracer	**86.5±1.6**	93.7±1.2
提案手法	**86.0±1.0**	94.6±0.4

る局所パターン集合の抽出という形で行い，検索要求時には 100 個程度の spatial codeword との単純な距離計算からなる識別用特徴量抽出を行うだけであるので，高速である．

3.4　局所的な属性構成要素の有効性

本節においては，局所的な属性構成要素の有効性を検証した．実験には AR DB 及び OMRON DB を利用した．また，検索時の学習には 100 枚の学習サンプルを用いた．提案手法は，オフラインで構築しておいた「局所的」基本パターン集合（spatial codebook）を用いて特徴抽出する．これと対立する概念として「入力領域全体」を用いた教師無学習による基本パターン抽出があるが，これに対する局所的基本パターンの効果を確認するため，そうした手法の代表的手法である NCC（Normalized Cross Correlation）及び PCA（Principal Component Analysis）[21] との比較を行った．結果は表 4 に示した（表中の ± の後の数値は 10 回のクロスバリデーションを行った際の標準偏差）．顔位置や顔向きが安定している AR DB においては，どの手法も精度に大差はない．一方で，大きな顔向き等が含まれる OMRON DB においては，顔全体を入力として処理を行う NCC や PCA を適用した結果に比べ，局所領域に注目する提案手法や Face Tracer の精度が高いことから，実用的には局所的な基本パターン抽出がこの種の問題に適していることが分かる．

3.5　MEMO 基準の有効性

2.1 において MEMO（maximum entropy and maximum orthogonality）基準に基づく codebook の構築方法を述べた．本節においては，この基準の有効性を実験により示す．実験においては spatial codeword の数を 50 に制限し，より各選択基準の差が出やすい状態にして ME 基準のみ及び MO 基準のみと MEMO 基準の比較実験を行った．結果は表 5 に示したとおりである（表中の ± の後の数値は 10 回のクロスバリデーションを行った際の標準偏差）．実験結果を見ると，ME 基準のみで選択した場合には，精度が大幅に低下している．これは各 spatial codeword が互いに独立にならず同じようなクラスタのみを表現する spatial codeword が選択されるからであると考えられる．また，MO 基準のみで選択したときにも，精度が低下している．これは codeword 間の独立性のみを重視しているので，ノイズに相当するような意味のないクラスタが選択されるからであると考えられる．これに対し両方の基準を組み合わせた MEMO 基準による結果は複数のデータセットにおいて他の基準より優れた結果となっており MEMO 基準が有効であることが分かる．したがって，ある程度の出現確率があり，かつ直行なクラスタをバランス良く選び出す MEMO 基準の有効性が実証された．

表 5　MEMO 基準の有効性
Table 5　Effectiveness of the MEMO criteria.

手法	OMRON DB	AR DB
ME	69.3±1.3	90.7±0.7
MO	78.7±1.4	90.1±0.5
MEMO	**79.8±0.8**	**94.1±0.5**

4.　考　　察

提案手法は多様な属性を認識するために，spatial codebook を用いて生じ得るパターンの集約を行っており，これは一般物体認識のフレームワークに類似である．一般物体認識の代表的フレームワークとして，画像内の特徴点検出（keypoint detection）により得られた局所的特徴を，特徴の生起する位置情報を使わず生起頻度のみで特徴量表現する Bag of Keypoints [9], [22]（BoK）が提案されている．位置情報を使わないのは，隠れが生じたり見え方が異なったりする対象に対して，いつも同じ特徴部位検出が行われる保証がないため，必ずしも同じ特徴領域間の比較が行えないからだと思われる．本研究では，顔位置は顔検出若しくは人体検出により位置決めがされており，局所領域同士の比較がある程度正確にできると想定しているので，位置情報を有効に使うことができたと考えられる．

5.　む　す　び

人物属性による人物検索は，増大する画像データの中から特定の特徴をもった人物を見つけるための重要な技術の一つである．一方，人物属性は非常に多様であるゆえ，それらすべてに設計時に対応することはほぼ不可能であり，必要に応じ新たな属性に対応しなが

ら検索を行うことが重要となる．このために少数の学習サンプルを用意するだけで即座に新たな属性を学習し高精度な検索が行える学習型人物属性検索のフレームワークを提案した．このためすべての処理を属性の学習時に行うのではなく，事前に属性を構成する局所的パターンを構築しておき，検索時には識別のための特徴抽出のみを行う手法を提案した．これによりすべての処理を同時に行う従来手法と比べ，効率的に高精度に検索が行えることを実験により示した．更に，特に学習サンプルが少ないときに提案手法が従来手法より精度が高いことも判明した．

今回の検討においては実験をしなかったが，例えば色情報を用いることにより髪色等にも応用が容易と考えられる．また様々な特徴量と組み合わせることにより更なる精度向上が得られると思われるが，それらは今後の課題である．更に，多くの属性は必ずしもメガネの有無等のように離散的な属性値をもつわけではなく，例えば髪型が短いものから長いものまで存在するように，連続的な属性値をもつ．筆者らが行った実験ではこのような場合にも，検索時の学習サンプルとして与えた画像と近い髪型の画像が検索されており，定性的にある程度効果的であることを確認したが，定量的な評価は困難であった．こうした連続値を有する属性に対する有効性の検証も今後の課題である．

謝辞　本研究の一部を支援して頂いた独立行政法人NEDO 次世代ロボット知能化技術開発プロジェクトに感謝します．

文　献

[1] B. Moghaddam and M.-H. Yang, “Learning gender with support faces,” IEEE Trans. Pattern Anal. Mach. Intell., vol.24, no.5, pp.707–711, 2002.

[2] S. Hosoi, E. Takikawa, and M. Kawade, “Ethnicity estimation with facial images,” Proc. FGR, p.195, 2004.

[3] J. Hayashi, H. Koshimizu, and S. Hata, “Age and gender estimation based on facial image analysis,” Knowledge-Based Intelligent Information and Engineering Systems, pp.863–869, 2003.

[4] X. Zhuang, X. Zhou, M. Johnson, and T. Huang, “Face age estimation using patch-based hidden Markov model supervectors,” Proc. ICPR, pp.1–4, 2008.

[5] M.J. Lyons, J. Budynek, A. Plantey, and S. Akamatsu, “Classifying facial attributes using a 2-d gabor wavelet and discriminant analysis,” Proc. FGR, p.202, 2000.

[6] T. Wilhelm, H.J. Bohme, and H.M. Gross, “Classification of face images for gender, age, facial expression, and identity,” Int’l Conf. on Artificial Neural Networks, pp.569–574, 2005.

[7] N. Kumar, P. Belhumeur, and S. Nayar, “FaceTracer: A search engine for large collections of images with faces,” Proc. Europican Conference on Computer Vision, pp.340–353, 2008.

[8] T. Leung and J. Malik, “Representing and recognizing the visual appearance of materials using three-dimensional textons,” IJCV, vol.43, pp.29–44, 2001.

[9] F. Li and P. Perona, “A bayesian hierarchical model for learning natural scene categories,” Proc. CVPR, pp.524–531, 2005.

[10] F. Jurie and B. Triggs, “Creating efficient codebooks for visual recognition,” Proc. ICCV, pp.604–610, 2005.

[11] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Lost in quantization: Improving particular object retrieval in large scale image databases,” Proc. CVPR, pp.1–8, 2008.

[12] J.C. vanGemert, J.-M. Geusebroek, and C.J. Veenman, “Kernel codebooks for scene categorization,” Proc. ECCV, pp.696–709, 2008.

[13] J. Wu and J.M. Rehg, “Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel,” Proc. ICCV, pp.630–637, 2009.

[14] A. Agarwal and B. Triggs, “Hyperfeatures — Multilevel local coding for visual recognition,” Proc. ECCV, vol.1, pp.30–43, 2006.

[15] A.M. Martinez and R. Benavente, “The AR face database,” CVC Tech. Report, p.202, June 1998.

[16] “The AR Face Database,” http://www.ece.osu.edu/~aleix/ARdatabase.html

[17] OMRON Corporation, “OKAO Vision,” http://www.omron.com/r_d/coretech/vision/okao.html

[18] C. Huang, H. Ai, Y. Li, and S. Lao, “High-performance rotation invariant multiview face detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.4, pp.671–686, April 2007.

[19] 木下航一，小西嘉典，勞　世紅，川出雅人，“3D モデル高速フィッティングによる顔特徴点検出・頭部姿勢推定,” MIRU2008 論文集，pp.1324–1329, 2008.

[20] T. Joachims, Learning to classify text using support vector machines, Kluwer, 2002.

[21] M. Turk and A. Pentland, “Eigenfaces for recognition,” J. Cognitive Neuroscience, vol.3, no.1, pp.71–96, 1991.

[22] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” ECCV Workshop on Statistical Learning in Computer Vision, pp.1–22, 2004.

（平成 22 年 3 月 23 日受付，6 月 5 日再受付）

井尻　善久　（学生員）

平 12 京都工繊大・工芸・機械システム卒，平 14 同大大学院修士課程了，同年オムロン（株）入社，現在に至る．平 21 より名大・情・メディア博士後期課程所属．人体検出・顔検出・顔認識等，人・顔画像処理の研究に従事．平 20 SSII デモオーディエンス賞受賞．平 21 SSII 高木賞受賞．平 22 PRMU 研究会研究奨励賞受賞．IEEE 会員．

勞　　世𤋮

1984 中国浙江大学電気工学科卒．昭 60〜平 4 京大・工・電気留学．同年オムロン（株）入社．以来人工知能，画像処理，パターン認識等の研究開発に従事．平 16 よりオムロン（株）コアテクノロジーセンタ技術専門職．平 21 SSII 高木賞受賞．IEEE 会員．

村瀬　　洋　（正員：フェロー）

昭 53 名大・工・電気卒．昭 55 同大大学院修士課程了．同年日本電信電話公社（現 NTT）入社．平 4 から 1 年間米国コロンビア大客員研究員．平 15 から名古屋大学大学院情報科学研究科教授，現在に至る．文字・図形認識，コンピュータビジョン，マルチメディア認識の研究に従事．工博．昭 60 本会学術奨励賞，平 6 IEEE-CVPR 最優秀論文賞，平 7 情報処理学会山下記念研究賞，平 8 IEEE-ICRA 最優秀ビデオ賞，平 13 高柳記念奨励賞，平 13 本会ソサイエティ論文賞，平 14 本会業績賞，平 15 文部科学大臣賞，平 16 IEEE Trans. MM 論文賞，ほか受賞．IEEE フェロー，情報処理学会会員．