コンピュータビジョンと
イメージメディア 124－2
（2000．11．21）

形状遷移情報の学習を用いた時系列画像からの手指形状推定

浜田康志　島田伸敬　白井良明
大阪大学大学院工学研究科電子制御機械工学専攻
〒565-0871 大阪府吹田市山田丘2-1
E-mail:hamada@cv.mech.eng.osaka-u.ac.jp

本論文では2台のカメラにより取得されるシルエット画像を用いて手指形状推定を行なう．まず入力された2眼画像に対し左右別々に輪郭を用いた特徴の抽出，特徴ベクトルの作成，モデルとの差の計算を行なう．左右2枚の画像がどれくらい特徴的かを示す値として形状の複雑度を計算し，この値を用いて左右の差を統合し全体の差をもとめる．そして全体の差が最小のモデル画像を選出する．また認識時の計算量を減らすために遷移ネットワークを用いる．まず学習段階で学習画像シーケンスから典型的な手指形状モデルを追加し，モデル間の可能な遷移を自動的に登録してネットワークを構築する．次に認識段階で遷移ネットワークよりマッチング候補を絞り込み効率良くマッチングを行なう．実際にいくつかの時系列画像シーケンスを用いて学習を行ないネットワークを使ったマッチングの効果を示す．

Hand Shape Estimation Using Image Transition Network

Yasushi Hamada,Nobutaka Shimada and Yoshiaki Shirai
Dept.of Computer-Controlled Mechanical Systems, Osaka University
2-1 Yamadaoka,Suita,osaka 565-0871,Japan
E-mail:hamada@cv.mech.eng.osaka-u.ac.jp

This paper presents a method of hand posture estimation from silhouette images taken by two cameras. First, we extract each feature from the left and right silhouette contour, make feature vectors using eigenspace method, and then compute distances between the input and the model. We define shape complexity for each image to evaluate how well the shape feature is represented in order to compute the total matching distance combining the left and right distance. For rapid search for the best-matched model image to the input, we use a transition network. In the offline learning phase, nodes and links of the network are automatically created by showing sample sequences. In the online shape estimation, we limit the matching candidates using the transition network. We show experiments of building the transition network and matching using the network.

1　はじめに

画像を用いたヒューマンインタフェースや手話理解がマウスやキーボードといった従来の入力装置にかわるものとしてに注目されてきている．以前から手指形状を調べる手法としてデーグローブ等のセンサを内蔵した器具を装着することがあったが，画像を用いることにより手の動作に制約を与えることなく手指形状を調べることができる．

最近の画像ベースの手指形状認識手法は2種類に分類される．

1つめは与えられた3次元モデルから可能な姿勢を生成し入力画像に最も合う姿勢を探索する方法である [1][2][3]. これらの手法は任意の姿勢の認識に効果的であるが多大な計算コストかかる.

2つめは学習シーケンスにおける画像や画像の特徴を登録し, 入力シーケンスを登録されたシーケンスに分類するものである [4][5][6]. この手法は限られた手指形状の認識のためだけに役立つモデルを登録するだけなので3次元形状を具体的に表さないが, 3次元形状を推定しないので計算量は少ない.

この論文では学習で登録されたモデル画像の中から最適なモデル画像を選ぶことにより手指形状を推定する手法を提案する. まず, 2眼画像の各画像について形状を推定するために手領域の輪郭を抽出し, これを元にした特徴ベクトルを算出する. そしていろいろな手指形状の画像をモデル画像として集め, 各モデル画像から特徴ベクトルを算出し, データベースに蓄える.

モデルとのマッチングでは, 入力画像と各モデル画像の差として特徴ベクトルの固有空間上での距離を求める. 2眼画像を用いているので右同士, 左同士を比較し2つの距離を求める.

このとき, 形状の特徴がどれくらい現れているかを示す形状の複雑度を左右各画像について求める. そして左右画像それぞれにおける特徴ベクトルの差を統合するために, 左右それぞれの差を形状の複雑度をもとに組み合わせ, 2眼画像全体の差を計算する. この差が最小のモデル画像を最適な手指形状モデルとする.

登録ジェスチャを増やしたとき, マッチングの対象となる手指形状モデルの数が多くなるため計算時間が増加する.

ジェスチャや手話を考える場合, 手指形状画像は時系列画像のシーケンスとして入力される. 時系列画像を用いた認識において, 3次元モデルを用いる手法は各関節の角度や運動の制約をもちいて探索範囲を限定することができるので効果的である. しかしこのような限定を画像の特徴を用いる手法で利用するのは簡単ではない.

そこで, この論文では手指形状変化の拘束によってマッチング候補を限定する手法を提案する. まず, 手指形状モデルと可能な遷移をノードとリンクによって表したネットワークを学習によって構築する. 認識時にはこのネットワークを用いて, シーケンスにおける次フレームでの形状を現フレームのモデルから遷移可能なモデルにのみ限定する. この遷移ネットワークの学習は処理に時間が掛かるのでオフラインで行ない, いくつかの学習シーケンスを示すことによって自動的にノードとリンクを作成する. また, あるシーケンスから獲得されるノードと良く似たノードを持つ別のシーケンスがあるとそれらのノードはマージされ, それらのジェスチャはマージされたノードを通じて遷移が可能となる.

実際にいくつかの学習画像シーケンスを用いて遷移ネットワークを構築し, 学習に用いたシーケンスではない新たなシーケンスを用いて認識を行なう. またマッチングの回数の比較を行ない計算量を減少させられることを示す.

2 特徴抽出

2.1 輪郭特徴

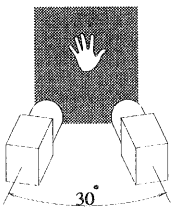


図 1: カメラ配置

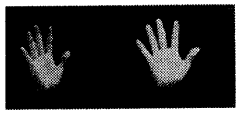


図 2: 2眼画像の例

手領域を簡単に獲得するために手領域は背景や服よりも明るいものとする.

(a) 取得画像　　(b) 輪郭特徴の抽出

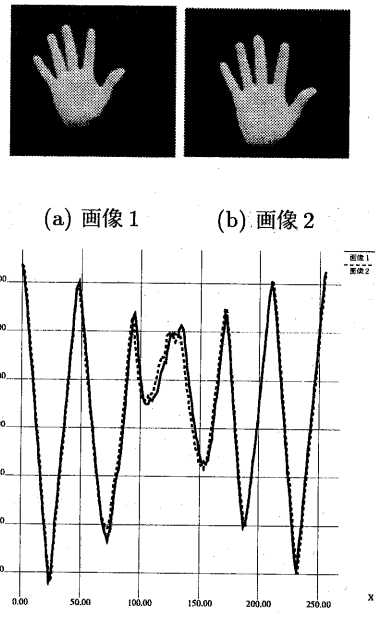
図 3: 特徴抽出

実験者の前方に横に並べて固定された2台のカメラにより左右2枚1組の手画像が獲得される (図1, 図2). 各画像に対して手領域を抽出し面積 S と重心 G を計算する. そして図3のように手領域の境界線上に等間隔に256個の点 $P_i(i = 1, \cdots, 256)$ をとり重心 G から各点までの距離 d_i を求め次式により大きさを正規化する.

$$r_i = \frac{d_i}{\sqrt{S}} \qquad (1)$$

この計算より形状特徴は位置や大きさの変化に対して不変になる.

またこの形状特徴は回転に依存するので配列の始点を変化させて要素を並べ直すことで回転を正規化する. もっとも重要な点として r_i の最大値, 最小値を始点にとり, ならびかえた $\boldsymbol{x} = \{r_1, \cdots, r_{256}\}^T$ を特徴ベクトルとする. 図4(a)(b) の二つの良く似た手指形状画像の特徴ベクトルを図4(c) に示す.



(a) 画像1　　(b) 画像2

(c) 良く似た特徴ベクトルの例

図 4: 良く似た画像の特徴ベクトル

2.2 固有空間の作成

学習段階では, 様々な手指形状をモデルとして登録する. まず, 効率よくマッチングを行なうために, 各モデル画像の特徴ベクトルより固有空間を構築する. 固有空間の基底は主成分分析によって得られる k 個の固有ベクトル $\boldsymbol{E} = [e_1, \cdots, e_k]$ とする. そして, この固有空間に投影されたモデル n の特徴ベクトル

$$\boldsymbol{g}_n = \boldsymbol{E}^T(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(n = 1, \cdots, M) \qquad (2)$$

をデータベースに蓄える.

認識段階では入力画像に対して大きさの正規化された距離 $\{r_i\}$ を同様に求める. 回転を正規化するために $\{r_i\}$ の最大値, 最小値を始点候補とするが, ロバストな正規化を行なうために極大値の大きい順, 極小値の小さい順に L 個の候補を求めそれぞれについてモデルとの比較を行なう. j番目の候補 $\boldsymbol{y}_j = \{r_{j,1}, \cdots, r_{j,256}\}^T$ は $\{r_i\}$ を並べ変えて生成される.

次に各 $\boldsymbol{y}_j$ を固有空間へ投影し圧縮された

特徴ベクトル

$$\boldsymbol{h}_j = \boldsymbol{E}^T(\boldsymbol{y}_j - \bar{\boldsymbol{x}}) \quad (3)$$

を計算する.

3 2眼画像を用いたマッチング

入力画像とモデル画像 n のマッチングは, 入力画像から作成される L 個の $\boldsymbol{h}_j$ とモデル画像 n の $\boldsymbol{g}_n$ を用い,

$$d_n = \min_{j=1,\cdots,L}(\|\boldsymbol{h}_j - \boldsymbol{g}_n\|) \quad (4)$$

により計算される. 入力に対する最適なモデルとの差は式4を用いてすべてのモデル画像との差を求め

$$d = \min_{n=1,\cdots,m}(d_n) \quad (5)$$

により計算される.

単眼画像ではマッチングを行なうのに有効な手指形状が獲得されないことがある. そこでこの章では2眼画像を用いたマッチングの手法について述べる.

3.1 形状複雑度を基にしたマッチング

入力された2眼画像の左右それぞれに対して式4を適用し, それぞれの差を計算する.

最適なモデルを推定するための左右それぞれの差の組み合わせ方として, 左右2つの差の平均を使う方法がある. しかしこの方法ではマッチングに有効でない側のシルエットによる影響を受けてしまう.

そこで, より複雑な形状ほど3次元手指形状を表すのに有効であると仮定し, 形状特徴の複雑度

$$c = \sum \frac{r_{i+k} - r_i}{k} \quad (6)$$

を定義する. k は実験的に求めた定数でここでは10としている.

2眼画像の左右それぞれに複雑度による重み w_l, w_r をかけ, 2つの差の重み平均によって最適モデルの推定を行なうこととする. ただし, もし一方の画像の複雑度が他方より明らかに大きいなら, 後者はマッチングに悪影響を及ぼすと考えて前者のみを使うこととする.

左右の複雑度 c_l, c_r によって以下のように場合分けを行ない重み付けの計算を行なう.

- もし $\sqrt{a_l^2 + a_r^2} \leq t_1$ または $\sqrt{a_l^2 + a_r^2} > t_1$, $\frac{1}{t_2} \leq \frac{a_l}{a_r} \leq t_2$ なら $w_l = c_l, w_r = c_r$ とする.
- もし $\sqrt{a_l^2 + a_r^2} > t_1$, $\frac{a_l}{a_r} < \frac{1}{t_2}, t_2 < \frac{a_l}{a_r}$ なら $w_l = 1, w_r = 0 (c_l > c_r$ のとき) とする.

t_1, t_2 は実験的に求めた定数である.

3.2 実験結果

まず典型的な手指形状の特徴ベクトルを用いて固有空間を構築する. 実験において固有空間の次元を徐々に増やしたとき12次元で認識率が飽和したので以降では12次元固有空間を用いる.

260枚のモデル画像と74枚の入力画像をもちいた認識実験では95.9%の認識率が得られた. 図5に入力と認識結果を示す.

4 遷移ネットワーク

ジェスチャや手話の認識において時系列画像のシーケンスが入力される. 時系列画像シーケンスではフレーム間で変化可能な手指形状を限定できる.

遷移ネットワークは手指形状モデルを表すノードと手指形状モデル間の可能な遷移を表すリンクにより構成される.

すべての手指形状の変化を学習させるのは多大な時間と努力を必要とするため完全な遷移ネットワークの構築は難しい. そこでこの

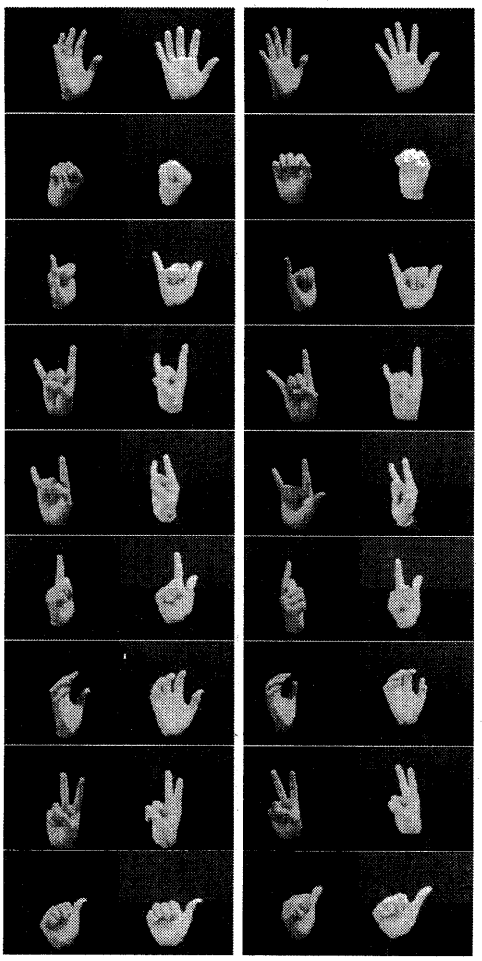


図 5: 認識結果

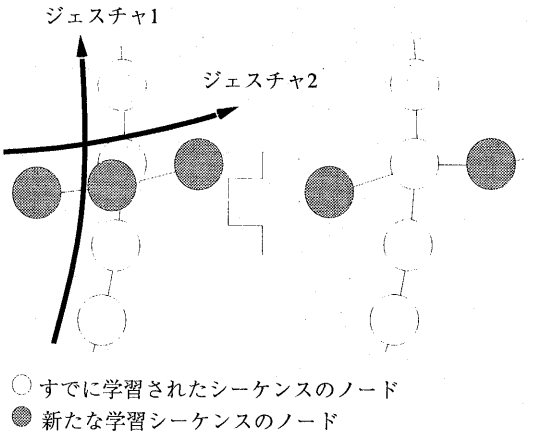


図 6: ノードのマージ

章では遷移ネットワークの学習の手法について述べ，いくつかの学習画像シーケンスを用いて遷移ネットワークを構築し認識実験を行う．

4.1 遷移ネットワークの構築

遷移ネットワークの学習段階では，学習画像シーケンスを与え，自動的に新しいノードおよびリンクを追加しネットワークを構築する．

あらかじめ典型的な手指形状モデルを登録しておく．学習画像シーケンスの各フレームに対して特徴ベクトルを求め，固有空間上に投影する．そして3章で述べたマッチングを行ないすでに登録されている中から最も似ているモデルの選出を行なう．

もし入力とモデルの差 d がしきい値よりも小さければ入力と選出されたモデルは同一形状をしているとみなし，このモデルのノードにマージする．(図 6) この時，この特徴ベクトルにつながるリンクもこのモデルのノードに連結する．

もし d がしきい値よりも大きければ一致する手指形状モデルはないとする．

次にどのモデルとも一致しなかったフレームを新たなモデルとして登録する． 1つのシーケンスにおいてモデル A に一致したフレームとモデル B に一致したフレームの間にどのモデルとも一致しなかったフレームが存在する場合を考える．(図 7) この時，どのモデルとも一致しなかったフレームの中でモデル A,B との差がもっとも大きいフレーム (図 7の C) を新たなモデルのノードとして遷移ネットワークに登録する． 新しく登録したモデルと他の一致するモデルのなかったフ

レームを比較し，その差がしきい値以下ならば同一形状をしているとみなす．この操作を新たな登録が為されなくなるまで繰り返す．

この処理をすべての学習画像シーケンスに対して行なうことで典型的な手指形状を表すノードおよびノード間の可能な遷移を示すリンクを登録し，遷移ネットワークを構築する．

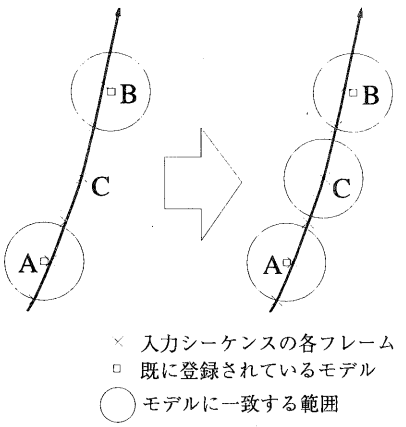


図 7: 新たなモデルの作成

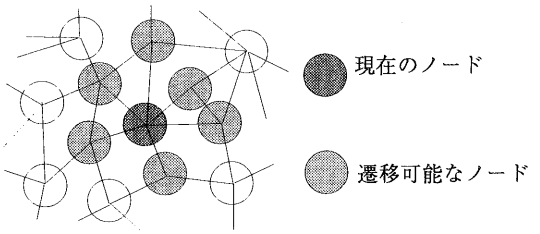
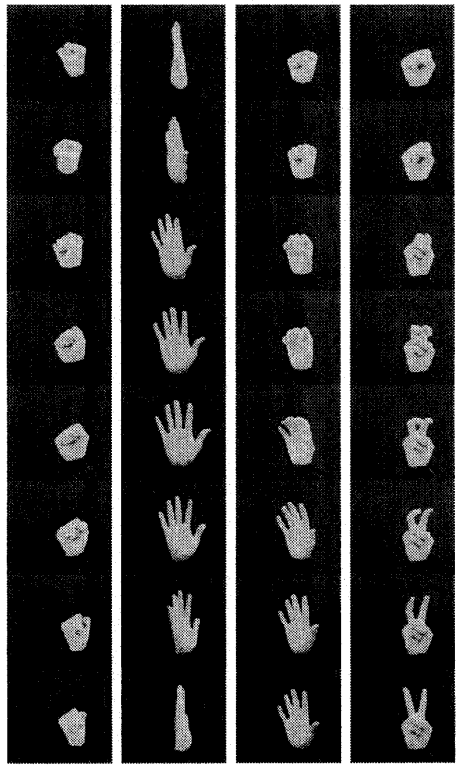


図 8: マッチング候補の限定

4.2 遷移ネットワークを用いた手指形状シーケンスの認識

入力画像シーケンスを認識する際，遷移ネットワークを次フレームの形状候補を限定するのに利用する．(図8) 現在のフレームにおけるマッチングによりある手指形状モデルのノードが選ばれると，次のフレームにおける形状モデル候補は現在のノードから遷移可能なノードに限定できる．この候補とのみマッチングを行なうことで計算コストを減少させる．



(a) ジェスチャ A　(b) ジェスチャ B　(c) ジェスチャ C　(d) ジェスチャ D

図 9: 学習に用いたジェスチャシーケンス

学習段階で2つ以上のジェスチャに含まれるノードがいくつか存在する．したがって，2つ以上のジェスチャを組み合わせた入力シーケンスに対し，このノードを通じて遷移可能なので認識することができる．例えば図6では学習段階でジェスチャ1と2の結合ノードが生じる．もし認識段階においてジェスチャ1と2が部分的に組み合わさったジェスチャが入力された場合にも，この遷移ネットワークを用いてうまく追跡を行なうことができる．

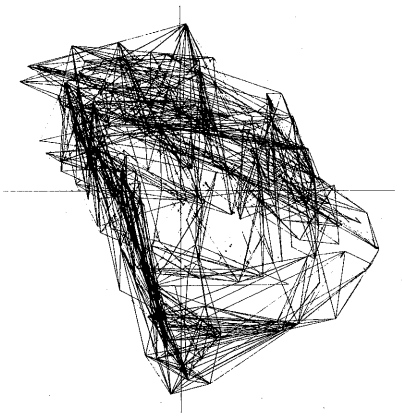


図 10: 遷移ネットワーク

5 実験結果

遷移ネットワークを構築し，ジェスチャシーケンスを認識する実験を行なった．

学習段階において 8 種類のジェスチャを用い，各ジェスチャに対し 5 から 35 個のシーケンスを用意した．各シーケンスの長さは 200 フレームとした．図 9に 8 種類の内の 4 種類のジェスチャを示す．

全体で 141 シーケンス 28200 フレームの画像から，258 個のノードを持つ遷移ネットワークが生成された．多くの良く似たフレームが一つのノードにマージされたため，ノードの総数を抑えることができた．図 10に生成されたネットワークを示す．図の 2 つの軸は固有空間での各ノードの 12 個の成分のうちのもっとも主要な 2 つの成分である．頂点と線はノードとリンクを表している．

認識段階では新たなジェスチャシーケンスを用いて実験をおこなった．図 12(a) はこのシーケンスが図 9 の学習ジェスチャ B から学習ジェスチャ C，そして学習ジェスチャ D へと変化することを示している．

この新しいシーケンスに対する認識結果を図 12(b) に示す．また，図 11に遷移ネットワークにおけるジェスチャ A-D を表すノードとリンク (灰色) と入力シーケンス (黒線) を示す．

この手法による計算コスト軽減の効果を調べるためにマッチング回数を比較した．従来手法における画像の認識では 258 枚のモデルすべてと比較するので，マッチング回数は 1 フレームあたり 258 回となる．したがって 200 フレームの認識を行なうためには平均 51600 回のマッチングが必要となる．

遷移ネットワークを用いる手法ではリンクされたノードの数が一定でないためマッチング回数は入力されたシーケンスにより変化するが，本章の実験に用いた入力シーケンスの場合，マッチング回数は 5789 回で従来手法の 11% であった．

現在，遷移ネットワークの 1 ノードあたりのリンクの数は平均 8.9 である．これはランダムシーケンスにおいてマッチングの回数は従来の手法に比べ約 3% になることを意味する．

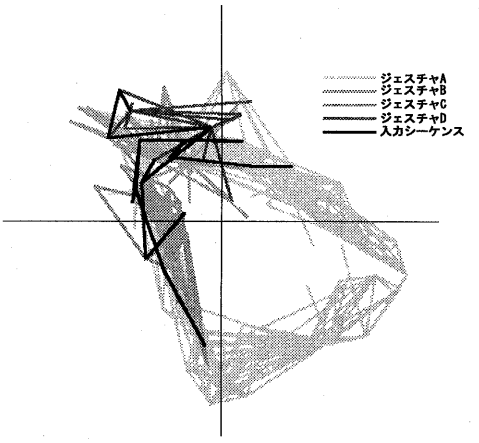


図 11: 入力シーケンスと入力シーケンスに含まれるジェスチャ

6 おわりに

2 台のカメラにより取得される輪郭画像を用いた手指形状推定の手法を述べた．学習段階では画像特徴の固有空間を用いてモデル画像データベースを作成した．また認識段階では左右の複雑度を評価し複雑度をもとにした

左右のマッチング結果の統合により最適なモデルを推定した.

次に, 時系列画像シーケンスの学習, 認識で手指形状遷移ネットワークを用いる手法を提案した. まず学習段階で学習画像シーケンスから遷移ネットワークを自動的に作成した. 認識段階では現フレームでの最適なノードから遷移可能なノードを形状候補として選びこの候補の中から次フレームでの最適なモデルを選出することで計算コストを削減した.

本論文で提案した手法は時系列画像の入力に対して効率良く手指形状を認識することができた. 今後の課題は手指形状のシーケンスを意味を持ったシーケンスとして認識することである.

参考文献

[1] N.Shimada, Y.Shirai, Y.Kuno and J.Miura: "Hand Gesture Estimation and Model Refinement using Monocular Camera". International Conference on Automatic Face and Gesture Recognition,pp.268-273,1998.

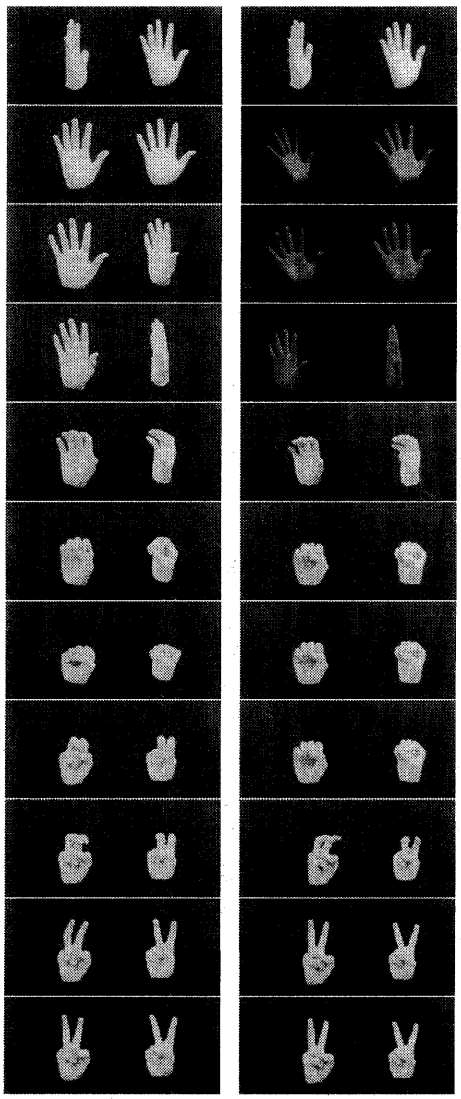
[2] 亀田 能成, 美濃 導彦, 池田 克夫: "シルエット画像からの関節物体の姿勢推定法". 電子情報通信学会論文誌,D-II Vol.J79-D-II,No.1,pp.26-35,1996.

[3] 岩井 儀雄, 八木 康史, 谷内田 正彦: "単眼動画像からの手の 3 次元運動と位置の推定". 電子情報通信学会論文誌,D-II Vol.J80-D-II,No.1,pp.44-55,1997.

[4] 木村 光佑, 島田 伸敬, 白井 良明: "CG 検索に基づく単眼シルエット画像を用いた三次元手指姿勢の推定". MIRU2000,pp.II145-II150,2000.

[5] M.J.Black and A.D.Jepson: "Eigen-Tracking: Robust Matching and Tracking of Articlated Objects Using a View-Based Representation". International Journal of Computer Vision,26(1),63-84,1998.

[6] T.Ahmad,C.J.Taylor,A.Lanitis,T.F.Cootes: "Tracking and recognising hand gestures, using statistical shape model". Image and Vision Computing,15,pp.345-352,1997.



(a) 入力　　(b) 認識結果

図 12: 遷移ネットワークを用いた認識結果