Proceedings of the VI Brazilian Conference on Neural Networks - VI Congresso Brasileiro de Redes Neurais
pp. 307–312, June 2–5, 2003 – Centro Universitário da FEI, São Paulo - SP - Brazil

Reconhecimento de Posturas Manuais Usando Redes Neurais

Marcus V. Lamar[1], Md. Shoaib Bhuiyan[2] e Akira Iwata[3]
[1]Departamento de Engenharia Elétrica
Universidade Federal do Paraná – UFPR
[2]Information Processing Center
Suzuka University of Medical Science and Technology
[3]Department of Electrical and Computer Engineering
Nagoya Institute of Technology - Japan
E-mails:lamar@eletrica.ufpr.br, iwata@elcom.nitech.ac.jp

Abstract

In this work we present a fast and efficient feature extraction method to be applied in visual gesture recognition systems. The method is based upon the using of Principal Component Analysis (PCA) to extract morphological information about 2D regions. The system performs the modeling of hand postures from color gloved video images. The performance of the feature extraction method is evaluated in applications of Japanese and American finger spelling automatic recognition system. The use of color gloves allows a fast tracking and complex hand model be extracted against natural backgrounds. A feedforward multiplayer perceptron neural network classifier achieved 89.4% of correct recognition rate in a set of 42 Japanese fingerspelling postures, and 94.5% in a 26 hand postures set extracted from American fingerspelling.

1. Introdução

O reconhecimento de gestos humanos é um campo de estudos bastante promissor, onde se busca uma interface mais natural entre homens e máquinas. O uso de dispositivos especiais, tais como, luvas instrumentalizadas tem atraído a atenção dos pesquisadores [1]. Tais sistemas atingem elevado grau de precisão no reconhecimento dos gestos e posturas manuais com baixa complexidade computacional, no entanto ainda são sistemas extremamente caros. Técnicas baseadas em visão computacional têm gradativamente chamado a atenção dos pesquisadores. Mesmo sendo técnicas com elevada complexidade computacional, requerendo processadores velozes para serem aplicadas em tempo real, devido ao barateamento do custo das câmeras digitais e ao crescente aumento do poder de processamento dos microprocessadores, aliados a técnicas ágeis como as apresentadas neste trabalho, têm se tornado técnicas populares.

A linguagem dos sinais é sem dúvida o mais complexo e bem estruturado conjunto de gestos humanos. O reconhecimento de gestos através do uso de visão estéreo [2] ou técnicas de visão 3D [3] são muito interessantes e podem produzir excelentes resultados, porém requerem alta carga computacional além de necessitarem de dispositivos mais caros que técnicas 2D usando somente uma simples e barata câmera colorida. Em pesquisas sobre a Linguagem Americana dos Sinais (*American Sign Language* - ASL), é conhecido que as posturas manuais não carregam muita informação para a identificação de um gesto executado [3,4]. Porém na linguagem dos sinais japonesa (*Japanese Sign Language* - JSL), o uso preciso das posturas manuais é de grande importância para a diferenciação dos diferentes gestos [5]. Então, em sistemas de tradução automática da JSL é muito importante o uso de técnicas rápidas e precisas para a extração e modelagem das mudanças temporais das posturas manuais a partir de um sinal de vídeo.

Este trabalho propõe uma nova abordagem para sistemas de reconhecimento de gestos. Primeiramente, apresentamos um método de extração de características para um sistema de visão 2D baseados em uma simples luva colorida, usado para modelar a postura e posição da mão a partir de um sinal de vídeo. A seguir propomos o uso de redes neurais para o reconhecimento dos gestos previamente segmentados no tempo. Apresentamos aqui, uma modelagem utilizando redes neurais independente do contexto e aplicamos a um sistema de reconhecimento de gestos independente da gramática. A Figura 1 mostra uma visão geral do sistema.

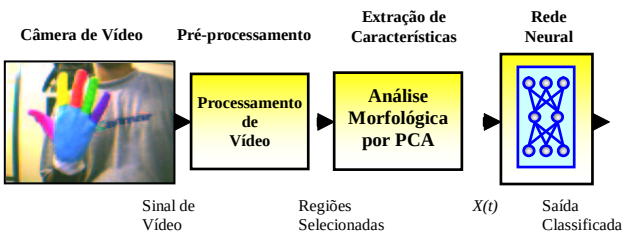


Figura 1: Visão Geral do Sistema

2. Modelagem da Mão a partir da Imagem

A localização da mão e extração da postura executada a partir de uma seqüência de imagens naturais são as duas mais complexas e importantes tarefas no reconhecimento de sinais manuais.

Neste trabalho utilizamos uma luva barata de pano, com a palma e os dedos pintados com diferentes cores, mostrada na Figura 1. Com esta luva, obtemos um rápido rastreamento e localização da mão nas imagens e relativamente complexa descrição do modelo da mão. É conhecido que não é necessário um preciso modelo 3-D da mão para aplicações de reconhecimento em tempo real das linguagens dos sinais[4]. Assim, luvas precisamente rotuladas, como apresentada em [6], onde a junção de cada dedo é pintada marcada com um anel de cor diferente, não são necessárias. Em [7] Iwai *et al*, propõe o uso de uma luva onde cada dedo é pintado com 2 cores, possibilitando a correta localização das pontas dos dedos. Mostramos neste trabalho que mesmo esta abordagem não é necessária. Estas luvas precisamente marcadas tendem a ser dependentes do usuário, pois dependem de aspectos anatômicos da mão, tais como o tamanho dos dedos.

Neste trabalho, cada quadro do sinal de vídeo é segmentado e as regiões de interesse, pertencentes à luva colorida, são detectadas utilizando as técnicas de processamento de vídeo técnicas propostas em [8].

2.1. Análise de Componentes Principais

Em reconhecimento de padrões é usual pré-processarmos os dados para a extrair somente as informações relevantes, de modo a reduzir a complexidade do projeto do classificador.

Neste trabalho buscamos extrair as características relevantes para a modelagem da mão a partir de uma imagem. Definimos um vetor de características $\vec{V}_c$ para cada região previamente selecionada R_c relativa a cor c, aplicando a Análise de Componentes Principais (*Principal Component Analysis* - PCA) à distribuição de localização dos pixels para cada região conexa.

Dada uma distribuição 2D X_c, com elementos $\vec{x}_c$ definidos por

$$\vec{x}_c = (i, j) \mid P(i, j) \in R_c \tag{1}$$

onde R_c é a região selecionada de cor c e $P(i, j)$ os pixels, localizados nas coordenadas (i, j), que pertencem a esta região. A centróide de cada região pode ser definida como o valor esperado de $\vec{x}_c$ de acordo com

$$\vec{\mu}_c = E\{\vec{x}_c\} \tag{2}$$

A matriz de covariância $\mathbf{C}_c$ associada a distribuição X_c é definida por

$$\mathbf{C}_c = E\left\{(\vec{x}_c - \vec{\mu}_c)(\vec{x}_c - \vec{\mu}_c)^T\right\} \tag{3}$$

Então, para cada região de interesse R_c, resolvemos o seguinte sistema de equações

$$\det\left(\mathbf{C}_c - \lambda_{j,c} I\right) = 0 \tag{4}$$

onde $j = 0,1$. A solução deste sistema fornece os autovetores $\vec{e}_{j,c}$ associados aos autovalores $\lambda_{j,c}$ da matriz de covariância da distribuição X_c. Com base nesta análise definimos o vetor de características para cada região através dos seguintes componentes:

a) Centróide Normalizada Relativa

A centróide normalizada relativa $\vec{\tau}_c$ caracteriza a posição relativa de um dedo em relação aos outros. Esta característica é invariante a deslocamentos paralelos e perpendiculares ao plano da câmera. Sua definição é

$$\vec{\tau}_c = \frac{\vec{\mu}_c - \frac{1}{k}\sum_{i=1}^{k}\vec{\mu}_i}{\max_{j=1\ldots k}\left\{\left|\vec{\mu}_j - \frac{1}{k}\sum_{i=1}^{k}\vec{\mu}_i\right|\right\}} \tag{5}$$

onde k é o número de cores de interesse detectadas na imagem. O conjunto de vetores $\vec{\tau}$ mapeia a posição dos dedos em um sistema de referências não inercial definido pela média das centróides $\vec{\mu}_c$, cujas componentes são limitadas ao intervalo [-1,1].

b) Razão dos Autovalores

Da teoria do PCA, é conhecido que os autovalores $\lambda_{j,c}$, fornece um senso sobre o espalhamento da distribuição X_c sobre os eixos principais definidos pelos autovetores $\vec{e}_{j,c}$. Usamos esta característica para descrever o estado de um dedo. A razão dos autovalores ζ_c é definida como

$$\zeta_c = 2 \cdot \frac{\lambda_{n,c}}{\lambda_{m,c}} - 1$$
$$\lambda_{m,c}, \lambda_{n,c} \in \lambda_{j,c} \,\big|\, \lambda_{m,c} \le \lambda_{n,c} \tag{6}$$

onde $\lambda_{m,c}$ e $\lambda_{n,c}$ são respectivamente o maior e o menor autovalor calculados a partir da Eq.(4).

Se ζ_c é um valor próximo a -1, significa que há um eixo principal na região R_c, podemos supor então que o dedo correspondente esteja esticado. Por outro lado, se ζ_c é próximo a 1, não há um eixo principal, supomos então que o dedo esteja curvado.

c) Direção Principal

Os autovetores $\vec{e}_{j,c}$ da matriz de covariância determinam os eixos principais da distribuição X_c. Deste modo define-se a direção principal normalizada θ_c como

$$\theta_c = \frac{2}{\pi}\arctan\left(\frac{e^x_{m,c}}{e^y_{m,c}}\right) - 1 \qquad (7)$$

onde $\vec{e}_{m,c} = e^x_{m,c}\cdot\vec{i} + e^y_{m,c}\cdot\vec{j}$ é o autovetor associado ao maior autovalor $\lambda_{m,c}$. O ângulo θ_c medido a partir da horizontal, define a orientação do dedo em um espaço normalizado 2D.

Cada região colorida R_c é caracterizada por um vetor de dimensão 4, definido por

$$\vec{V}_c = \left(\tau^x_c, \tau^y_c, \zeta_c, \theta_c\right) \qquad (8)$$

onde τ^x_c e τ^y_c são as componentes do vetor centróide normalizada $\vec{\tau}_c$.

Para determinar a postura da mão, caracterizada pelas posturas dos 5 dedos, necessitamos de um vetor característica de dimensão 20.

A influência da inclusão ou não da informação correspondente à palma da mão no vetor de característica foi analisado em [8]. Devido a problemas de oclusão sofrido pelos dedos, a palma da mão é rastreada e sua centróide usada apenas para definir a localização da mão $\vec{P}_h = (i, j)$ na imagem. Para manter a informação de deslocamento da mão independente da distância entre o usuário e a câmera, definimos o vetor direção do movimento $\vec{P}_v(t)$ em um instante de tempo t como

$$\vec{P}_v(t) = \frac{\vec{P}_h(t) - \vec{P}_h(t-1)}{\left|\vec{P}_h(t) - \vec{P}_h(t-1)\right|} = \left(P^x_v(t), P^y_v(t)\right) \qquad (9)$$

A informação contida em $\vec{P}_v(t)$ é bastante ruidosa em posturas estáticas. Este problema é superado pela aplicação do algoritmo de detecção de movimento, usado para segmentar o sinal de vídeo.

Deste modo, um vetor dependente do tempo $\vec{V}(t)$ de dimensão 22 é gerado para caracterizar a postura da mão independente da sua localização e a informação da trajetória, definido como

$$\vec{V}(t) = \left(\vec{P}_v(t), \vec{V}_1(t), \vec{V}_2(t), \vec{V}_3(t), \vec{V}_4(t), \vec{V}_5(t)\right) \qquad (10)$$

3. Resultados e discussões

Esta seção apresenta os resultados dos experimentos realizados usando uma estação Silicon Graphics, 130MHz, com uma câmera de vídeo colorida e placa de aquisição Vino, usando quadros de tamanho 160×120 pixels e 24 bits/pixel.

A segmentação temporal realizada neste trabalho visa descartar os quadros que possuam movimentos transitórios entre duas posturas estáticas. A análise da quantidade de movimento de uma cena está baseada na energia de 2 consecutivas diferenças entre quadros, de modo a podermos descartar quadros transitórios e considerarmos apenas quadros estáticos. Definimos a energia da diferença entre 2 quadros como

$$E_d(f_i, f_{i-1}) = \sum_{l=1}^{nl}\sum_{c}^{nc}\left(f_i(l,c) - f_{i-1}(l,c)\right)^2 \qquad (11)$$

onde $f_i(l,c)$ é o valor do pixel da posição (*linha,coluna*) no i-ésimo quadro. Dada a condição

$$E_d(f_i, f_{i-1}) \geq E_{th1} \qquad (12)$$

onde E_{th1} é um valor de limiar definido e fixo. Se a energia da diferença entre quadros satisfizer esta condição, o quadro é classificado como dinâmico, sendo então descartado. Se a condição não for satisfeita, o quadro é considerado estático e deve ser processado. Para evitar que uma mesma postura seja considerada mais de uma vez, o sistema espera por um novo quadro estático somente após a mão se movimentar novamente. Isto é analisado por outro limiar de valor mais elevado E_{th2}. Podemos pensar esta abordagem como a aplicação de um laço de histerese.

Aplicando este procedimento, pode-se processar os quadros em tempo real, conseguindo-se um pequeno atraso de 0.5 segundos quando um quadro estático é analisado pelo método de extração das características seguido da rede neural.

3.1. Alfabeto Manual Japonês

No alfabeto manual japonês, apresentado na Figura 2, 41 posturas manuais estáticas e 5 movimentos das mãos são usadas para representar os 46 caracteres *kana* do alfabeto japonês. Considerou-se a representação da letra "も" (mo) como sendo uma

postura estática, uma vez que sua postura final é completamente diferente das demais. Assim temos um total de 42 diferentes posturas manuais que serão objeto de estudo deste trabalho.

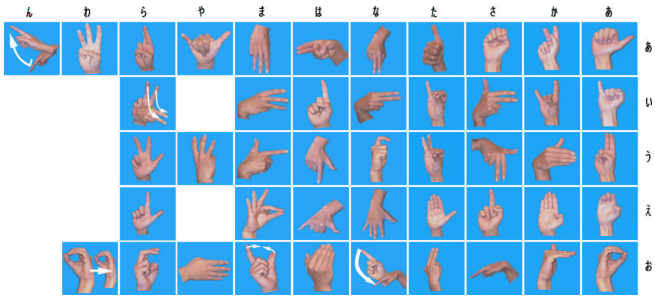


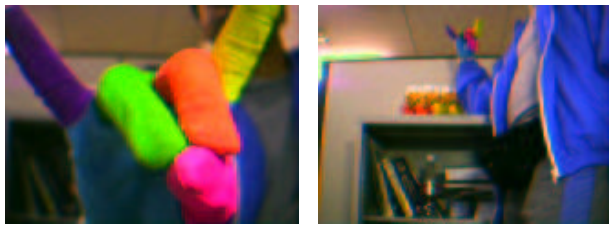
Figura 2: Alfabeto Manual Japonês

O método de extração de características proposto foi aplicado, gerando um vetor de dimensão 20 para cada imagem, descartando a informação direção do movimento $\vec{P}_v(t)$. Cada uma das 42 posturas foi executada 30 vezes por um usuário em ocasiões diferentes para gerar o conjunto de treino. O conjunto de teste foi gerado pelo mesmo usuário, alterando-se as condições ambientais,tais como iluminação, fundo e roupas. Os conjuntos de treino e teste possuem o mesmo número de vetores, 1260 vetores de dimensão 20. Nenhum cuidado especial foi tomado quanto as condições ambientais e roupas, devido ao fato que o sistema apresentado é suposto ser robusto o suficiente para trabalhar em condições normais.

Para este experimento, treinamos uma rede MLP usando o algoritmo de *Backpropagation*. A estrutura dotada é composta de 3 camadas., Tendo 20 unidades na camada de entrada, correspondente as 20 dimensões do vetor característica de entrada. A camada intermediária é composta de 42 neurônios. A camada de saída também é composta de 42 neurônios relativos às 42 diferentes posturas manuais a serem classificados da JSL. Usou-se a função sigmóide como função de ativação para todos os neurônios.

A oclusão é um problema comum encontrado em reconhecimento de gestos utilizando técnicas visuais. Neste trabalho, mapeamos uma região colorida oculta em um vetor nulo, e então deixamos a rede neural aprender esta característica.

Os resultados obtidos indicam uma taxa de acertos de 89.4% para as 42 posturas do conjunto de teste.A Figura 3 mostra dois exemplos extremos da capacidade de modelagem da postura manual pelo método proposto, permitindo o correto reconhecimento da mesmo com diferentes distâncias da câmera e fundos naturais.



(a) (b)

Figura 3: Exemplo de posturas manuais corretamente reconhecidas para a letra “き”(ki).

3.2. Alfabeto Manual Americano

Para analisar o desempenho do método de extração de características apresentado, em um mais amplamente conhecido conjunto de posturas manuais, treinamos outra rede neural para o reconhecimento do alfabeto manual da Linguagem Americana dos Sinais (ASL), mostrado na Figura 4.

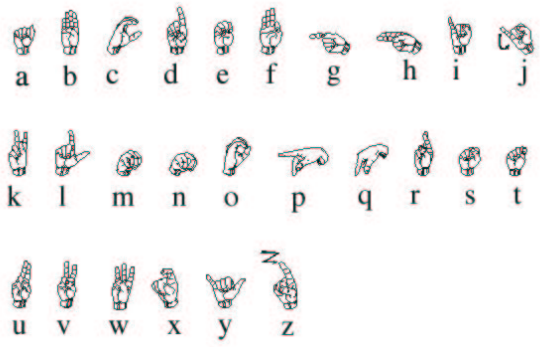


Figura 4: Alfabeto Manual Americano

No alfabeto manual da ASL, as 26 letras são representadas por 24 posturas estáticas e 2 movimentos manuais. Os movimentos são utilizados na representação das letras “J” e “Z”. Neste trabalho, considerou-se apenas a postura final da mão para estas duas letras, considerando o movimento existente como sendo parte dos movimentos transitórios, que são descartados pelo sistema de segmentação temporal. Novamente aqui, esta estratégia é eficiente, uma vez que a postura final para estas letras são bastante distintas das demais. O banco de dados foi gerado por um usuário, que executou cada uma das posturas 60 vezes em dias e condições ambientais diferentes. O conjunto de treino é composto por 30 amostras de cada postura, sendo as restantes 30 imagens reservadas para compor o conjunto de teste, gerando um total de 780 imagens por conjunto.

Foi treinada uma rede neural MLP de 3 camadas, contendo 20 unidades de entrada, 26 neurônios na camada intermediária e também 26 neurônios na camada de saída, representando os 26 os caracteres. O sistema atingiu uma taxa de reconhecimento de 94.5% para o conjunto de teste.

A Tabela 1 apresenta os piores desempenhos obtidos para cada alfabeto manual, e suas taxas de reconhecimento e erro.

Tabela 1: Análise de erros para os piores casos

JSL			ASL		
Letra	Correto	Erro	Letra	Correto	Erro
“す”(su)	46.7%	53.3%	C	70.0%	30.0%
“ろ”(ro)	53.3%	46.7%	T	73.3%	26.7%
“こ”(ko)	56.7%	43.3%	M	76.7%	23.3%
“さ”(sa)	63.3%	36.7%	Q	76.7%	23.3%

A análise destes piores resultados indica que a rede neural teve mais dificuldades em classificar as posturas que apresentam dedos ocultos, forçando a rede a analisar vetores com grande número de componentes de valor zero.

3.3. Comparação com trabalhos publicados

A Tabela 2 sumariza uma comparação entre o desempenho do método apresentado, isto é, extração de características seguida de redes neurais, e outros métodos publicados. A primeira coluna apresenta o autor do método e a segunda coluna mostra o tipo de sistema de classificação utilizado. A terceira coluna o ambiente utilizado, isto é, dispositivos auxiliares e tipo de fundo para os métodos baseados em visão computacional, onde F.H. é usado para designar Fundo Homogêneo e F.C. Fundo Complexo. A quarta coluna apresenta o número de classes do problema enfrentado e a finalmente taxa de reconhecimento atingida.

Tabela 2: Comparação com trabalhos publicados

Autor	Classificador	Ambiente	Classes	Taxa
Iwai	Árvore de decisão	Luva Colorida F.H.	26	92.3%
Murakami	Redes Neurais Recorrentes	Data Glove	42	98.0%
Triesch	Casamento de Grafos	Mão Livre F.C.	10	86.2%
Birk	Classificador de Bayes	Mão Livre F.H.	25	99.7%
Lamar	Redes Neurais	Luva Colorida F.C.	42	89.4%
Lamar	Redes Neurais	Luva Colorida F.C.	26	94.5%

Iwai *et al* [7] apresentou um sistema baseado em árvore de decisão, utilizando luvas coloridas para reconhecer 26 posturas manuais em um ambiente bastante controlado, com fundo homogêneo negro, e considerando uma imagem específica da mão, conseguindo uma taxa de reconhecimento de 92.3%. Murakami e Taguchi [9] usaram uma luva instrumentalizada (*data glove*) comercial em um sistema dependente do usuário, com classificação por redes neurais e reportaram uma taxa de 98.0% de correto reconhecimento para um conjunto de 42 posturas manuais. Triesch *et al*[10] usaram casamento de grafos elásticos para classificar 10 posturas manuais sem o uso de qualquer tipo de luva e com fundos complexos, reportando 86.2% de correta classificação. Birk *et al.* [11] relatam 99.7% de correto reconhecimento de 25 sinais da ASL usando PCA e classificador de Bayes em um sistema sem luvas, dependente do usuário, com fundo homogêneo negro e ainda utilizando uma imagem em *close* da mão. Comparando estes resultados publicados com os resultados atingidos neste trabalho, podemos notar que o sistema proposto representa um bom compromisso entre taxa de reconhecimento e a flexibilid ade do sinal de vídeo de entrada, por não requerer fundo homogêneo e executar, de maneira rápida e precisa, a localização da mão na imagem.

4. Conclusões

Este trabalho apresentou um método de extração de características que utiliza análise de componentes principais para extrair informações morfológicas de uma região 2-D a partir de um sinal de vídeo digital. Aplicou-se este método para modelar posturas manuais utilizando uma luva colorida, que permite uma rápida localização da mão nas imagens. Analisamos o desempenho das técnicas apresentadas em um sistema de reconhecimento automático do alfabeto manual baseado em redes neurais. A automática segmentação temporal do sinal de vídeo de entrada é realizada e uma imagem estática é selecionada para análise, gerando um vetor de características de dimensão 20. Experimento com 42 diferentes posturas manuais do alfabeto manual japonês foi realizado, atingindo uma taxa de reconhecimento de 89.4%. Treinamos também o sistema para reconhecer o as 26 posturas do alfabeto manual da ASL, atingindo 94.5% de taxa reconhecimento. Nossos esforços estão atualmente dedicados a gerar um sistema capaz de reconhecer as 26 posturas do alfabeto manual da Linguagem Brasileira dos Sinais (LIBRAS).

Referências

[1] S. Tamura and S. Kawasaki, Recognition of Sign Language Motion Images, in *Pattern Recognition*, Vol.21, No 4, pp. 343-353, 1988.

[2] Q. Delamarre and O. Faugeras, Finding Pose of Hand in Video Images: A Stereo-Based Approach, in *Proceedings of 3th International Conference on Automatic Face and Gesture Recognition*, pp. 585-590, Nara, Japan, 1998.

[3] C. Vogler and D. Metaxas, ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis, in *Proceedings of the International Conference on Computer Vision*, pp. 363-369, Mumbai, India, 1998.

[4] T. Starner and A. Pentland, Visual Recognition of American Sign Language Using Hidden Markov Models, in *International Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, 1995.

[5] S. Tamura and S. Kawasaki, Recognition of Sign Language Motion Images, in *Pattern Recognition*, Vol. 21, N. 4, pp.343-353, 1988.

[6] B. Dorner, Hand Shape Identification and Tracking for Sign Language Interpretation, in *Workshop "Looking at People: Recognition and Interpretation of Human Action"*, IJCAI-93, Chambery, France, 1993.

[7] Y. Iwai, K. Watanabe, Y. Yagi, and M. Yachida, Gesture Recognition Using Colored Gloves, in *IEEE International Conference on Patter Recognition*, Vol. A, pp.662-666, Viena, 1996

[8] Lamar, M. V. *Hand Getsure Recognition using T-CombNET: A Neural Network dedicated to Temporal Information Processing*, Ph.D. Thesis, Nagoya Institute of Technology, Japan, 2001.

[9] K. Murakami and H. Taguchi, Gesture Recognition using Recurrent Neural Networks, in *Proc. Of CHI'91*, pp. 237-242, 1991.

[10] J. Triesch and C. Von de Malsburg, Robust Classification of Hand Postures against Complex Backgrounds, in *Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition*, Vermont, USA, 1996.

[11] H. Birk, T. B. Moeslund, and C. B. Madsen, Real-Time Recognition of Hand Alphabet Gestures using Principal Components Analysis, in *Proc. 10th Scandinavian Conf. on Image Analysis*, Lappenranta, Finland, 1997.