# UMA COMPARAÇÃO ENTRE REDES NEURAIS ARTIFICIAS E MÁQUINAS DE VETORES DE SUPORTE PARA RECONHECIMENTO DE POSTURAS MANUAIS EM TEMPO-REAL

TICIANO A. C. BRAGATTO[1], GABRIEL I. S. RUAS[1], MARCUS V. LAMAR[2]

*[1]Departamento de Engenharia Elétrica – Universidade de Brasília*
*Campus Universitário Darcy Ribeiro*
*Caixa Postal 4386 – CEP 70.910-900*
*Brasília – DF*

*[2]Departamento de Ciência da Computação – Universidade de Brasília*
*Campus Universitário Darcy Ribeiro*
*Caixa Postal 4466 – CEP 70.910-900*
*Brasília – DF*

bragatto@unb.br, gruas@unb.br, lamar@unb.br

**Abstract:** This paper presents a comparison between different artificial intelligence techniques used for finger spelling recognition on a real-time system, with emphasis on Support Vector Machines and Artificial Neural Networks but also using other well-known methods. The comparison is based on the computational complexity, correct rate and time taken to process the required information. The tests are based in two models of ANN and a binary tree structured multi-class SVM. The results show that SVM has an equivalent recognition rate as ANN and it is computationally cheaper and faster than ANN.

**Keywords:** Support Vector Machines, Artificial Neural Networks, Finger Spelling

**Resumo:** Este artigo apresenta uma comparação entre diferentes técnicas de inteligência artificial usadas para reconhecimento de posturas manuais em tempo-real, com ênfase em Máquinas de Vetores de Suporte e Redes Neurais Artificiais, mas também utilizando outros métodos conhecidos. A comparação é baseada na complexidade computacional, taxa de acerto e tempo necessário para processar a informação. Os testes estão baseados em duas implementações de RNAs e SVM multi-classes estruturadas em árvore binária. Os resultados mostram que SVMs obtém uma taxa de reconhecimento equivalente às RNAs, com menor complexidade computacional.

**Palavras-chave:** Máquinas de Vetores de Suporte, Redes Neurais Artificiais, Classificação de Posturas Manuais

## 1. Introdução

Hoje em dia, a comunicação entre homem e máquina é feita principalmente usando dispositivos como teclado e *mouse*, que não são formas naturais de comunicação para humanos. Portanto, pesquisas que permitam ao computador ver e ouvir são muito importantes para a criação de um sistema que seja capaz de interagir com seres humanos de forma mais natural. Reconhecimento de voz é um campo muito popular de estudos, com muitos trabalhos já publicados. Atualmente, com o aumento de capacidade de processamento dos computadores pessoais, o reconhecimento de gestos está começando a atrair a atenção de muitos pesquisadores. No entanto, o desenvolvimento de sistemas de processamento de vídeo digital em tempo-real ainda é uma tarefa desafiadora em máquinas de baixo custo e baixo poder de processamento.

A língua dos sinais, usada por deficientes auditivos, apresenta o mais complexo e gramaticalmente estruturado conjunto de gestos manuais utilizado pelo homem [Tamura 1989]. O processamento de sinais de vídeo contendo este tipo de linguagem e seu reconhecimento e tradução automática é um campo de estudos ainda pouco explorado.

Neste trabalho, realizamos uma comparação e apresentamos os resultados obtidos pelo uso de duas técnicas de classificação no problema do reconhecimento automático de posturas manuais. Redes neurais artificiais (RNA) e máquinas de vetores de suporte (*Support Vector Machines* – SVM) têm se mostrado técnicas concorrentes em diversas aplicações. Em processamento de vídeo em tempo-real, além da qualidade da classificação, a complexidade computacional do classificador é um ponto importante a ser considerado. O custo computacional e a eficiência na classificação das RNA e SVM são analisados e comparadas com métodos clássicos de classificação, tais como K-NN (*K-Nearest Neighbors*) e *Template Matching* (TM).

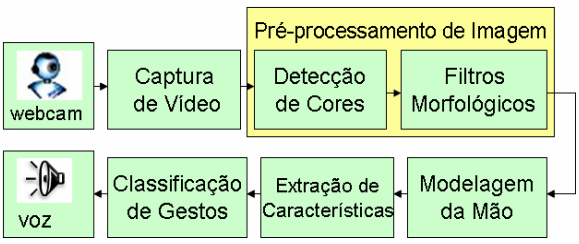


Figura 1. Diagrama em blocos do sistema de reconhecimento de posturas manuais

## 2. Sistema de Reconhecimento de Gestos Manuais

O diagrama de blocos do sistema de processamento de vídeo em tempo-real utilizado é apresentado na Fig. 1. Este sistema está em desenvolvimento e visa reconhecer gestos de duas mãos, isto é, localizar, seguir e analisar a posição das mãos em um espaço 2-D juntamente com a análise temporal da variação das posturas manuais [Lamar 2000].

Este trabalho está focado no estudo de um método eficiente para a classificação em tempo-real da postura manual já segmentada. No entanto, para completo entendimento do problema, torna-se necessário realizar uma breve explanação do sistema onde o classificador está inserido.

A fim de diminuir o esforço computacional da etapa de localização e rastreamento das mãos, usamos uma luva colorida, onde cada dedo possui uma cor diferente, apresentada na Fig. 2. Tal método foi testado e aprovado por diversos pesquisadores [Abe 2000][Bray 2002][Starner 1995]. O uso de uma luva colorida permite uma detecção e modelagem mais rápida das posturas dos dedos do que sistemas que utilizem imagens da mão livre, que necessitam um estágio complexo de detecção de cor de pele e segmentação dos dedos [Wysoski 2002]. Cada módulo da Fig. 1 é brevemente explicado a seguir.

Figura 2. A luva colorida usada neste trabalho

### *2.1 Estágio de captura*

O sistema está sendo desenvolvido em ambiente Windows® , usando a linguagem C em conjunto com a biblioteca de código aberto para visão computacional da Intel, OpenCV [Sudra 2002]. O OpenCV usa o Direct Show, do DirectX para efetuar a captura, permitindo a aquisição de vídeo em taxas altas com uma simples webcam USB.

### *2.2. Pré-Processamento da Imagem*

A cada quadro capturado, o estágio de pré-processamento é executado, afim localizar a mão na imagem. O pré-processamento é composto por um classificador de cores baseado em RNA [Bragatto 2005] seguido por filtros morfológicos resultando em 6 regiões conexas de pixels que pertencem a cada um dos 5 dedos e da palma da mão.

### *2.3. Modelo da Postura Manual*

As regiões conexas previamente detectadas são modeladas de acordo com o modelo proposto por Lamar et. Al [Lamar 2000], que extrai de cada região conexa um vetor de 4 dimensões a fim de caracterizar sua posição relativa e formato bidimensional aproximado por uma elipsóide. Usando somente as informações dos dedos, um vetor de 20 dimensões, variante no tempo é extraído e usado para representar a postura manual a cada quadro da seqüência de vídeo. Deste modo, um gesto manual é caracterizado por uma sequência temporal de vetores de dimensão 22-D, onde duas componentes correspondem a centróide da palma da mão a fim de localizá-la na imagem, e as 20 dimensões restantes correspondem ao vetor modelo da postura.

A Fig. 3 mostra as 26 posturas utilizadas do alfabeto da Língua Brasileira dos Sinais(LIBRAS). Deve-se notar que as letras J, K, X e Z usam movimento da mão em sua representação. Para classificar estes movimentos como somente posturas estáticas, considera-se somente a postura final de cada gesto.

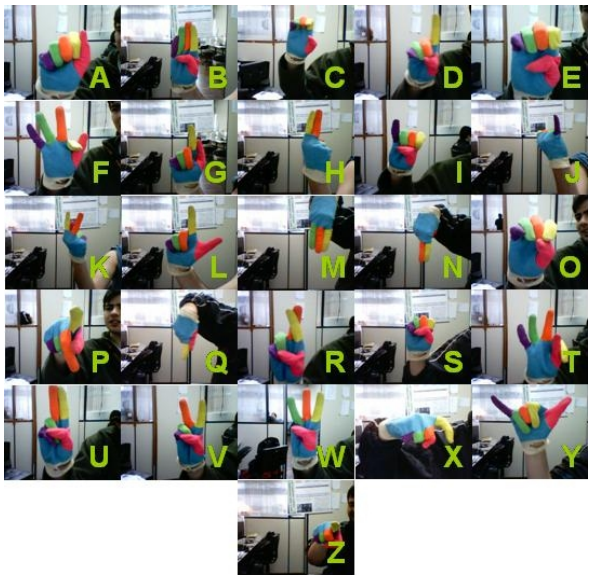


Figura 3. Alfabeto de Posturas em LIBRAS

### *2.4. Classificação da Postura Manual*

Como o foco principal deste trabalho é a análise das posturas estáticas da mão, a informação sobre a localização da mesma na imagem não é necessária. O sistema classifica continuamente apenas a postura manual a cada quadro da seqüência de vídeo, obtendo deste modo uma descrição de como esta postura evolui ao longo do tempo.

O estágio que necessita maior esforço computacional é o pré-processamento da imagem. Todo pixel do quadro deve ser analisado, classificado e as regiões conexas devem ser também processadas para gerar o vetor de características de cada postura ma-

nual em uma sequência de vídeo. No entanto, o classificador de saída impacta de forma também significativa no desempenho geral do sistema em tempo-real, pois a postura é classificada em todos os quadros.

## 3. Redes Neurais Artificiais de Baixa Complexidade Computacional

Para permitir a implementação de sistemas que demandem de grande capacidade de computação, tal como processamento em tempo-real de vídeo baseados em redes neurais, estudos de complexidade computacional de implementações de tais sistemas têm sido feitos por muito pesquisadores [Anderson 1988][Hertz 1991][Shimada 2006]. A maioria visa a construção de arquiteturas computacionalmente eficientes, considerando o número de neurônios e camadas, bem como o tipo de função de ativação de cada neurônio. Na implementação de sistemas que usem RNAs do tipo *MultiLayer Perceptron* (MLP), a função de ativação é o item mais custoso computacionalmente, pois envolve o cálculo de uma função não-linear. Uma das funções mais amplamente usadas em redes MLP é a função sigmoidal, que pode ser definida por

$$\Phi(x) = \frac{1}{1+e^{-x}} \qquad (1)$$

onde $x$ é a soma ponderada das entradas do neurônio.

A função sigmoidal requer o cálculo de uma exponenciação, uma soma e uma divisão. A fim de reduzir esta complexidade, Bragatto et al. [Bragatto 2005] propôs o uso de uma aproximação segmentalmente linear para a implementação desta função de ativação em sistema de tempo-real. Para uma aproximação em três linhas, a função é aproximada por

$$\Phi(x) = \begin{cases} 0 & x < -2.2 \\ 1 & x > 2.2 \\ 0.2273x + 0{,}5 & -2.2 \le x \le 2.2 \end{cases} \qquad (2)$$

resultando em uma função composta por apenas uma multiplicação, uma soma e duas comparações. A figura 3 mostra a função sigmoidal e sua aproximação por três retas.

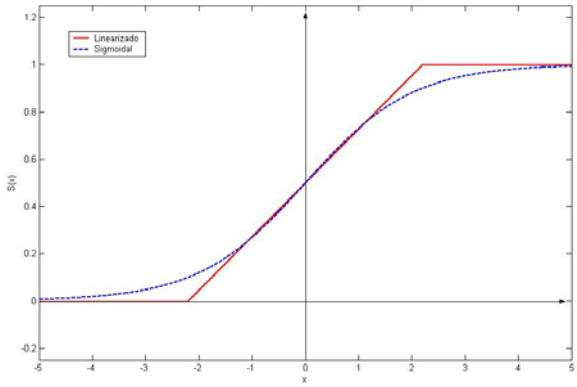

Figura 3. Funções de Ativação

Um índice de complexidade computacional (ICC) é definido para comparar o ganho de velocidade usando a aproximação. O ICC é um índice baseado na atribuição de pesos relativos para cada operação matemática, lógica e instruções, usado para analisar uma seção de código de programa [Bragatto 2005]. O ICC é usado neste trabalho para comparar resultados obtidos.

## 4. Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (*Support Vector Machines* – SVM) [Vapnik 1998] são um conjunto de métodos relacionados de aprendizado supervisionado usados para classificação e regressão. O fundamento das SVMs é fazer o aumento de dimensão do problema, até que seja linearmente separável por um hiperplano.

As SVM foram originalmente concebidas para tratar de maneira ótima problemas de classificação binária, isto é, em apenas duas classes. Para SVM multi-classes[Weston 1998], conhecidas como k-classes, SVMs do tipo *one–against–the–rest* são usadas amplamente. No problema da classificação da postura manual em soletração na língua dos sinais, temos um total de 26 classes (letras), sendo então necessário o uso de 26 SVMs, gerando uma elevada complexidade computacional.

Para testar e usar um sistema com essa característica, é necessário executar as 26 SVMs e decidir entre elas qual dos resultados positivo seria correto, uma vez que pode-se obter mais de uma resposta positiva.

O método utilizado neste trabalho atua como uma árvore de decisão binária[Fei 2006]. Se o resultado de uma SVM for 1, então a classe deve ser considerada correta, caso contrário, o sistema executa a próxima SVM, que usará uma classe a menos para seu treinamento, uma vez que a classe anterior já foi desconsiderada.

Um primeiro resultado desta metodologia é a necessidade de apenas (k − 1) SVMs para solução de um problema de k classes. Um segundo resultado é que, se a primeira SVM apresentar um resultado positivo, o sistema considera esta classe como correta e

não executa mais nenhuma SVM, reduzindo assim a complexidade computacional total do método.

Para aplicar esta técnica eficientemente deve-se considerar cuidadosamente a ordem de avaliação das SVMs. Considerou-se neste trabalho, que quanto maior a probabilidade de ocorrência de uma classe (letra), mais acima da árvore binária ela deve ficar, poupando esforço computacional.

Sabe-se que a soletração usando alfabeto manual é utilizada no contexto da Língua Brasileira dos Sinais (LIBRAS) para descrever principalmente nomes próprios de pessoas e lugares. No entanto, neste trabalho, a ordem selecionada das letras foi a probabilidade de ocorrência no português brasileiro escrito [Freqüência 2004], devido ao fato deste já ter sido foco de estudos estatísticos.

No cálculo do ICC as probabilidades de cada letra são consideradas, definindo uma média ponderada com os pesos sendo as probabilidades e os valores sendo o ICC de cada SVM (letra).

As 25 SVMs foram treinadas usando 3 tipos de kernel (linear, quadrático e polinomial), com o mesmo conjunto de treino das RNAs. As SVMs para cada letra foram selecionadas pela taxa de acerto, calculada usando o mesmo conjunto de testes das RNAs. Deste modo, cada uma das 25 SVMs possui diferentes kernels e um número diferente de vetores de suporte.

Além da redução da complexidade computacional, outra vantagem desta topologia de árvore de SVMs é a possibilidade de implementação de um módulo de treinamento dentro do próprio sistema. A topologia permite que se adicione uma nova postura manual ao classificador já treinado apenas inserindo o SVM treinado no topo da árvore, ao invés de treinar cada SVM novamente, tornando o sistema facilmente personalizável.

## 5. Resultados

Os resultados apresentados foram obtidos usando um microcomputador Pentium IV, 2.4 GHz, 512 Mbytes RAM e uma simples Webcam com imagens de 352x288 pixels, 24 bits/pixel e taxa de quadros de 30fps. O ambiente usado é de um típico laboratório de pesquisas em computação, com fundo complexo a sem controle de luminosidade, conforme apresentado na Fig. 2.

Os dados de treinamento, utilizados por todos os métodos implementados, consiste em um conjunto de imagens, formado por 26 posturas, com 25 repetições, de um mesmo usuário, tornando o sistema dependente do usuário. O conjunto de testes é composto por 5 imagens de cada uma das 26 posturas.

### *5.1. Redes Neurais de Baixa Complexidade*

Para estes testes foram implementadas 36 RNAs, com uma ou duas camadas escondidas e o número de neurônios variando de 30 a 200, com intervalo de 10 neurônios.

As duas redes escolhidas foram as menores e com melhores resultados para 1 e 2 camadas escondidas. A primeira (RNA1) é uma rede MLP com 1 camada escondida de 100 neurônios e a segunda (RNA2) contém duas camadas escondidas, cada uma com 30 neurônios.

A diferença de complexidade computacional destas redes pode ser analisada primeiramente pelo número de pesos de cada uma. A primeira rede, com 100 neurônios na camada escondida possui 4.726 pesos sinápticos, enquanto a segunda rede, mesmo com 2 camadas escondidas, possui 2.366 pesos, ficando evidente o menor custo computacional da segunda estrutura.

### *5.2. Máquinas de Vetores de Suporte*

A Tabela 1 mostra o ICC para o classificador de cada letra, bem como sua contribuição para o ICC do classificador geral da função e as probabilidades.

Tabela 1. Índices de Complexidade Computacional das SVMs

| Letra | ICC SVM | ICC Letra | Probabilidade | Contribuição |
|---|---|---|---|---|
| A | 494 | 495 | 14,63% | 72,42 |
| E | 494 | 989 | 12,57% | 124,32 |
| O | 1220 | 2209 | 10,73% | 237,03 |
| S | 535 | 2744 | 7,81% | 214,31 |
| R | 435 | 3179 | 6,53% | 207,60 |
| I | 617 | 3796 | 6,18% | 234,60 |
| N | 535 | 4331 | 5,05% | 218,72 |
| D | 658 | 4989 | 4,99% | 248,96 |
| M | 289 | 5278 | 4,74% | 250,19 |
| U | 371 | 5649 | 4,63% | 261,56 |
| T | 948 | 6597 | 4,34% | 286,32 |
| C | 1550 | 8147 | 3,88% | 316,11 |
| L | 494 | 8641 | 2,78% | 240,23 |
| P | 453 | 9094 | 2,52% | 229,18 |
| V | 453 | 9547 | 1,67% | 159,44 |
| G | 412 | 9959 | 1,30% | 129,47 |
| H | 453 | 10412 | 1,28% | 133,28 |
| Q | 412 | 10824 | 1,20% | 129,89 |
| B | 561 | 11385 | 1,04% | 118,41 |
| F | 647 | 12032 | 1,02% | 122,73 |
| Z | 412 | 12444 | 0,47% | 58,49 |
| J | 371 | 12815 | 0,40% | 51,26 |
| X | 289 | 13104 | 0,21% | 27,52 |
| K | 166 | 13270 | 0,02% | 2,65 |
| Y | 166 | 13436 | 0,01% | 1,34 |
| W | -- | 13437 | 0,01% | 1,34 |

O ICC do classificador geral é calculado pela soma dos ICCs dos classificadores ponderados pela probabilidade de ocorrência de cada letra, resultando em 4.077,37.

O ICC do classificador de cada letra é calculado pela soma do ICC de todas as SVMs de níveis superiores da árvore binária. Por exemplo, o ICC da letra S é calculado pela soma de 494(A), 494(E), 1.220(O), 535(S) e 1(pela atribuição).

As diferenças entre cada ICC são decorrentes dos diferentes tipos de kernel e do número de vetores de suporte.

A SVM mais complexa é a responsável pela identificação da letra "C", com 36 vetores de suporte e função de kernel polinomial.

### *5.3. Classificação das Posturas Manuais*

A Tabela 2 apresenta uma comparação do uso de RNA, SVM e classificadores clássicos k-NN (*k-Nearest Neighbors*) e TM (*Template Matching*), utilizando se em ambos 2 tipos diferentes de medidas de distâncias, a Distância Euclideana e *City Block*, no estágio de reconhecimento da postura manual.

Tabela 2. Resultados dos classificadores de postura manual

| Método de Classificação | ICC | Taxa de Acerto |
|---|---|---|
| TM – Euclideana | 1.560 | 84.6% |
| TM – *City Block* | 1.050 | 92.3% |
| k-NN – Euclideana | 31.210 | 93.8% |
| k-NN – *City Block* | 21.018 | 95.3% |
| RNA1 – Sigmoidal | 13.547 | 99.2% |
| RNA1 – Linearizado | 9.583 | 99.2% |
| RNA2 – Sigmoidal | 7.471 | 99.2% |
| RNA2 – Linearizado | 4.765 | 99.2% |
| SVM | 4.077 | 99.2% |

Neste experimento, RNA e SVM obtiveram taxas de reconhecimento superiores aos métodos clássicos. Devido ao ruído inerente aos sinais de entrada, k-NN, como é bem conhecido, não se apresenta como um método de classificação adequando para este tipo de sinal. RNAs e SVM são capazes de modelar melhor o problema devido a sua robustez ao ruído.

O uso da aproximação da função de ativação sigmoidal por seu equivalente segmentalmente linear com 3 retas, não apresentou impacto na qualidade do reconhecimento das redes neurais, atingindo 99.2%.

TM é o método mais leve computacionalmente, no entanto é o mais sensível à presença de ruído, gerando o pior desempenho de classificação. Comparando-se as redes RNA1 e RNA2, podemos concluir que o uso de 2 camadas escondidas é capaz de resolver o problema com menos esforço computacional, devido ao menor número de neurônios e pesos, e maior interconectividade entre os mesmos, quando comparado a uma rede neural com uma única camada escondida com mais neurônios. No entanto, SVM apresenta complexidade computacional ainda mais reduzida que RNA2, com desempenho similar.

O tempo de processamento é uma medida consistente da complexidade computacional de um algoritmo. O classificador de posturas é executado apenas uma vez por quadro, portanto seu impacto no desempenho geral do sistema é pequeno quando comparado com outros estágios. O ganho de velocidade obtido com o uso de rede neural RNA2 de baixa complexidade computacional sobre a mesma rede neural com função de ativação sigmoidal é de cerca de 1 quadro por segundo, aumentando a taxa de processamento de 27.1 para 28.1 quadros por segundo.

Com o uso de SVM estruturadas em árvore binária para a classificação, é atingida a taxa de 29.97 quadros por segundo, limite imposto pelo sistema de captura de vídeo utilizado.

## 6. Conclusões

Este artigo apresenta uma comparação do uso de SVM e redes neurais no contexto de um sistema de reconhecimento em tempo-real de posturas manuais aplicado a soletração na Língua Brasileira dos Sinais. Este trabalho faz o uso de funções de ativação linearmente segmentadas em duas redes neurais artificiais e SVM estruturadas em árvore binária, a fim de reduzir a complexidade computacional do classificador de posturas. Comparações com técnicas clássicas de classificação, k-NN e TM foram também realizadas. O uso de uma rede neural de baixa complexidade computacional permite uma considerável redução do tempo de processamento do sistema, atingindo taxas de 28.1 quadros por segundo. O uso de Máquinas de Vetores de Suporte, no entanto, apresentou-se como sendo mais econômico computacionalmente, atingindo a taxa limite do sistema de aquisição de vídeo, 29.97 quadros por segundo em um classificador contínuo do alfabeto manual da Linguagem Brasileira dos Sinais.

O sistema de classificação baseado em SVM implementado neste trabalho, mostra-se um classificador mais rápido que RNAs de baixa complexidade computacional, atingindo taxas de reconhecimento similares.

## Referências Bibliográficas


Abe, K., Saito, H., e Ozawa, S.. 3-d drawing system via hand motion recognition from two cameras. Systems, Man, and Cybernetics, 2000 IEEE International Conference on, 2(840-845), 2000.

Anderson, J. A.. A simple neural network generating an interactive memory. pages 181–192, 1988.

Bragatto, T. A. C., Sugawara, J. Y., Benso, V. A. P., e Lamar, M. V.. Reconhecimento de gestos em tempo-real utilizando uma rede neural artificial de baixa complexidade computacional para detecção de cores. Brazilian Symposium on

Computer Graphics and Image Processing, 2005.

Bray, M., Sidenbladh, H., e Eklundh, J.-O.. Recognition of gestures in the context of speech. Pattern Recognition, 2002. Proceedings. 16th International Conference on, 1(356-359), 2002.

Fei, B. e Liu, J.. Binary tree of SVM: a new fast multiclass training and classification algorithm. Neural Networks, IEEE Transactions on, Vol.17, Iss.3, 696- 704,2006.

Freqüência da ocorrência de letras no português. http://www.numaboa.com.br/criptologia/matematica, 2004.

Lamar, M. V., Bhuiyan, Md. S., e Iwata, A.. Hand gesture recognition using T-COMBNet: A new neural network model. IEICE Trans. on Information and Systems, E83-D(11):1986–1995, 2000.

Shimada, M., Iwasaki, S. e Asakura, T.. Finger spelling recognition using neural network with pattern recognition model. SICE 2003 Annual Conference, Vol.3, Iss., 4-6, 2458-2463, 2003.

Starner, T. e Pentland, A.. Visual recognition of american sign language using hidden markov models. International Workshop on Automatic Face and Gesture Recognition, 1995.

Sudra, G.. Seminar medizinische simulationssysteme intel opencv. http://wwwiaim.ira.uka.de/Teaching/SeminarMedizin/, 2002.

Vapnik, V.. Statistical Learning Theory. Wiley-Interscience, 1998.

Weston, J. e Watkins, C.. Multi-class support vector machines. Technical Report CSD–TR–98–04, 1998.

Wysoski, S. G., Lamar, M. V., Kuroyanagi, S., e Iwata, A.. A rotation invariant approach on static-gesture recognition using boundary histograms and neural networks. 9th International Conference on Neural Information Processing, (726- 729), 2002.