EDUARDO SCHNELL E SCHÜHLI

RECONHECIMENTO DE GESTOS DE MAESTRO UTILIZANDO REDES NEURAIS ARTIFICIAIS PARCIALMENTE RECORRENTES

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Engenharia Elétrica. Programa de Pós-Graduação em Engenharia Elétrica - PPGEE, Departamento de Engenharia Elétrica, Setor de Tecnologia, Universidade Federal do Paraná.

Orientador: Prof. Marcus Vinicius Lamar, Ph.D.
Co-orientador: Prof. Marcelo Wanderley Mortensen, Ph.D.

Curitiba

2005

RECONHECIMENTO DE GESTOS DE MAESTRO UTILIZANDO REDES NEURAIS ARTIFICIAIS PARCIALMENTE RECORRENTES

Eduardo Schnell e Schühli

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Paraná

_____

Prof. Marcus Vinicius Lamar, Ph.D.
Orientador

_____

Prof. Oscar da Costa Gouveia Filho, Dr.
Coordenador do Programa em Pós-Graduação em Engenharia Elétrica

Banca Examinadora

_____

Prof. Marcus Vinicius Lamar, Ph.D.
Presidente

_____

Ewaldo Luiz de Mattos Mehl, Dr.

_____

Prof. Wilson Arnaldo Artuzi Júnior, Ph.D.

_____

Leandro dos Santos Coelho, Dr.

O coração do prudente adquire o conhecimento, e o ouvido dos sábios busca a sabedoria.

Provébios 18:15

Agradecimentos

A minha família e amigos que me apoiaram no decorrer do trabalho, especialmente à minha amada esposa Gisele, que me amou e suportou todos os dias. Ao professor Marcus Vinícius Lamar (“Kiko”), que em momento algum negou auxilio e viabilizou a execução deste trabalho com seus vastos conhecimentos e sua inesgotável paciência. A Marcelo Mortensen e Paul Kolesnik pelo tempo despendido em auxílio e compartilhamento de material. A Deus, que me deu capacidade, motivação e oportunidade de finalizar este trabalho, a Ele seja dada toda a glória.

SUMÁRIO

Página

Lista de Abreviações e Acrônimos

HMM	*Hidden Markov Models*
ANN	*Artificial Neural Network*
DTW	*Dynamic Time Warping*
RNN	*Recurrent Neural Network*
MLP	*Multi Layer Perceptron*
LVQ	*Learning Vector Quantization*
TDNN	*Time Delay Neural Network*
VQ	*Vector Quantization*
T-CombNET	*Temporal CombNET*
BPM	Batidas Por Minuto
SOM	*Self Organizing Map*

Lista de Figuras

Página

Lista de Tabelas

Página

Resumo

Os gestos de um maestro têm o objetivo de guiar outros músicos de uma orquestra na execução de uma música. Tempo, dinâmica e expressão são algumas das responsabilidades do maestro na música. Para que os gestos possam ser identificados pelos outros músicos eles seguem uma gramática definida. Nesta pesquisa foram realizados testes utilizando métodos computacionais para fazer o reconhecimento de tempo, dinâmica e gestos de expressão de ambas as mãos. Tempo e dinâmica foram extraídos usando análise da velocidade vertical da mão. O sistema de reconhecimento de gestos de maestro foi implementado utilizando redes neurais parcialmente recorrentes do tipo Elman e T-CombNET após um estágio de processamento de imagem. Os resultados são comparados com os obtidos utilizando classificador baseado em HMM (*Hidden Markov Models*).

Palavras Chave

Reconhecimento de gestos, Redes Neurais, T-CombNET e Rede neural parcialmente recorrente de Elman.

Abstract

The conducting gestures have the goal to lead the musicians in a music execution. Tempo, dynamic and expression are some of the conducting responsibilities in the music. To be recognized by the musician, the gestures follow well defined grammatical rules.In this research, tests were done using computational methods to recognize the conducting gestures, as tempo, dynamic and expressive gestures from both hands. Taking advantage of image processing, tempo and dynamic were extracted using analysis of the vertical velocity from the right hand. The conducting gesture recognition system was implemented using Elman and T-CombNET neural network structure after an image processing stage and a comparison have been done with a system based on HMM (Hidden Markov Models).

Keywords

Gesture recognition, Neural Network, T-CombNET, Elman.

Capítulo 1

Introdução

O reconhecimento de gestos é um promissor campo de estudos na área de visão computacional. Muitos estudos têm sido desenvolvidos na busca de novas técnicas para compreender os gestos. As grandes semelhanças com o processamento de voz fizeram com que técnicas como *Hidden Markov Models* (HMM)[36] fossem utilizadas também no reconhecimento de gestos em sua fase inicial. A utilização de redes neurais artificiais tem obtido bons resultados [10], [30], [31], [32], [33], e o desenvolvimento de novas estruturas capazes de tratar naturalmente este problema é um amplo objeto de estudos.

O reconhecimento de gestos de maestros é estudado desde os anos 70 e inúmeros estudos foram desenvolvidos até hoje [1], [2], [3], [4], [5], [6], [7], [8]. Os resultados obtidos têm sido de grande valia no desenvolvimento de novas técnicas para a identificação de padrões.

O maestro pode ser definido como um músico que utiliza a expressão do seu corpo para guiar outros músicos na reprodução de uma música. Um maestro talentoso utiliza movimentos de todo o seu corpo para reger a orquestra, porém os gestos definidos são realizados por ambas as mãos. Faculdades de regência no mundo despenderam esforços na definição de uma gramática envolvendo os gestos básicos do maestro, compartilhada hoje pela maioria dos músicos. Uma boa definição desta gramática pode ser vista no estudo de Rudolph [18].

1.1 Classificação dos gestos

Os gestos de maestro podem ser separados por grupos, classificados pelo efeito que eles geram na música. A Figura 1-1 mostra a hierarquia da gramática básica.

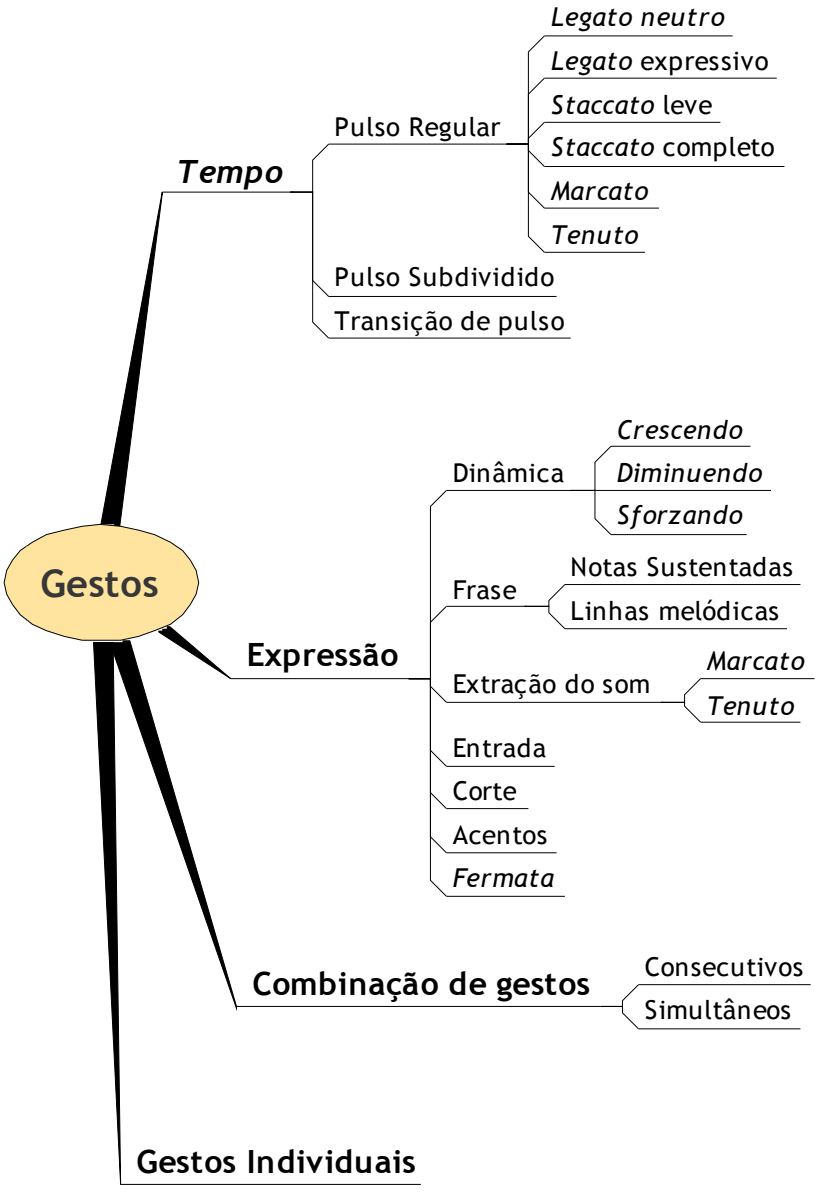


Figura 1-1: Hierarquia dos gestos na gramática básica de Rudolph [18].

Na execução dos movimentos, os gestos de tempo são principalmente executados pela mão direita, enquanto os gestos de expressão são executados pela mão esquerda. No entanto, essa divisão não é rígida, e alguns gestos de expressão podem ser executados pela mão direita. Os gestos de tempo podem ser representados por ambas as mãos em casos específicos, porém, sempre fazendo o mesmo movimento. Apenas em raríssimos casos é utilizada a mão esquerda para indicar tempo.

Neste trabalho, esta separação espacial dos movimentos foi utilizada para a identificação dos gestos. Assumiu-se que apenas a mão direita realiza os movimentos de tempo e ambas as mãos são responsáveis pelos gestos de expressão. Tentando abranger a maioria da gramática básica, utilizado-se um grupo de gestos composto por gestos de expressão e pulso regular.

1.2 Elementos musicais

Este trabalho foi desenvolvido com o objetivo de reconhecer a intervenção de um maestro na música, e por isso é importante destacar alguns elementos musicais envolvidos.

1.2.1 *Pulso (beat)*

O pulso é a unidade temporal da música. O grupo de pulsos e suas acentuações, repetido sequencialmente durante toda execução da música, ou parte dela, compõe um compasso.

1.2.2 *Tempo*

O *tempo* corresponde à velocidade em que os pulsos ocorrem. Como os pulsos são os marcos na música, o *tempo* interfere na velocidade em que a música é executada. A unidade convencional para *tempo* é batidas por minuto (pulsos por minuto). A percepção humana é capaz de distinguir o *tempo* até 240 batidas por minuto

1.2.3 *Dinâmica*

É a variação da amplitude em que o músico toca cada nota, influenciando na intensidade sonora da música.

1.2.4 *Articulação*

O elemento articulação identificado na pesquisa está relacionado com a maneira de tocar cada nota. O tocar da notas curtas e separadas é conhecido como *staccato* e o tocar das notas deixando soar, conectando-as é denominado *legato*.

1.2.5 *Expressão*

A expressão com que o músico deve tocar certo trecho da música também é indicada pelo maestro. Por exemplo, os momentos em que a música deve crescer em intensidade e parar instantaneamente é representado pelo *crescendo+corte*.

1.3 Revisão bibliográfica

O estudo dos gestos, suas gramáticas e a melhor forma de fazer o seu reconhecimento utilizando computadores têm sido objeto de muitos estudos.

O estudo dos gestos tem diversas aplicações. A mais comum delas é o reconhecimento de linguagem de sinais, tendo como público alvo as pessoas portadoras de deficiência auditiva.

Um dos primeiros estudos em reconhecimento de gestos foi realizado por Tamura e Kawasaki [30], fazendo o reconhecimento de linguagem de sinais japonesa utilizando *template matching*. Alguns estudos foram feitos utilizando HMM. Em 1991, Takahashi e Kishino [31] utilizaram redes neurais para o reconhecimento e Murakami e Taguchi [32] introduziram o uso de redes neurais recorrentes (RNN – *Recurrent Neural Network*).

Em 2001, Lamar e Iwata apresentaram bons resultados no reconhecimento de gestos utilizando a rede neural T-CombNET [10].

Utilizando os mesmos princípios do reconhecimento de gestos na linguagem de sinais, também muitas pesquisas foram desenvolvidas no reconhecimento de gestos de maestros [1], [2], [3], [4], [5], [6], [7], [8]. No entanto, no reconhecimento de gestos de maestros, diferentemente dos estudos na linguagem de sinais, na maioria dos casos a movimentação dos dedos não é importante, apenas a posição da mão é necessária para a identificação. Diferentes soluções foram usadas para a captura da posição da mão. A Tabela 1-1 descreve, em ordem cronológica, as diferentes soluções utilizadas para a captura do movimento.

A maioria dos estudos focaram na extração dos parâmetros de *tempo* e *dinâmica*, oriundos da mão direita do maestro.

O uso de redes neurais para a identificação dos gestos foi introduzido por Ilmonem e Takala [27] em 1999 e um ano antes Usa e Mochida [25] utilizaram HMM para o reconhecimento. Ambos os estudos também foram apenas nos gestos da mão direita.

Tabela 1-1: Soluções para captura do movimento das mãos de maestro.

Ano	Pesquisador	Entrada no sistema
1976	Mathews [19]	Teclado com botões e *joystick*.
1980	Buxton, Smith e Baecker [1]	Movimento 2D de um *mouse*.
1983	Haflich e Burns [5]	*Rangerfinder ultrasonic* de câmeras polaroid.
1989	Max Matthews [20]	Bastão, *joystick* e botões.
1989	Keane e Gross [6]	Bastão com elementos que se encostam na execução de um pulso e uma chave de pé.
1989	Morita e Hashimoto [21]	Câmera CCD e luvas brancas.
1990	Morita e Hashimoto [22]	Câmera CCD em conjunto com uma luva aquisitora de dados para a mão esquerda.
1991	Max Matthews [23]	Baqueta que emite ondas em rádio freqüência.
1996	Tobey e Fujinaga [24]	*Buchla Lightning batons*, 3D.
1998	Usa e Mochida [25]	Acelerômetros 2D, vídeo e sensor de respiração.
1998	Marrin e Picard [26]	Jaqueta com extensômetros, sensor de respiração, batimento cardíaco e de temperatura do corpo.
1999	Ilmonem e Takala [27]	Sensores magnéticos.
2000	Segen, Majumder e Gluckman [28]	2 câmeras.
2003	Murphy, Anderson e Jensen [29]	2 câmeras, frente e lado, compondo posição 3D.

Em 2004, Paul Kolesnik [8] percebendo a falta de pesquisas no estudo dos gestos da mão esquerda de maestro, incorporou na sua pesquisa o reconhecimento dos gestos de expressão realizados pela mão esquerda. Utilizando duas câmeras de vídeo de

baixo custo para a aquisição do movimento e HMM para o reconhecimento, Kolesnik obteve resultados significativos no reconhecimento da gramática básica.

1.4 Proposta

O presente trabalho teve como objetivo a utilização de uma nova metodologia para reconhecer os gestos do maestro. A identificação dos gestos de ambas as mãos diferenciou a pesquisa da maioria dos estudos realizados no assunto. As redes neurais tinham sido utilizadas por Ilmonem e Takala, e também Garnet, porém seus estudos foram apenas na identificação dos gestos da mão direita.

A utilização de redes neurais teve como motivação a comparação com os resultados obtidos com HMM por Kolesnik [8], tendo em vista que ambos os métodos são os mais utilizados para reconhecimento de padrões dependentes do tempo. A dificuldade encontrada no treinamento de redes neurais para reconhecimento de gestos motivou a utilização da rede neural T-CombNET, que apresenta solução para simplificar este problema [10]. A rede Elman [40] também foi utilizada nos testes graças a sua eficiência em resolver problemas temporais.

Este trabalho faz parte do projeto de uma orquestra digital, que vem sendo desenvolvido por diversos pesquisadores em conjunto com a Universidade McGill, no Canadá. O objetivo é desenvolver um sistema onde a única interação humana seja através do maestro, que regerá uma orquestra virtual.

1.5 Estrutura da dissertação

Esta dissertação está assim estruturada: No Capítulo 2 são descritos os conceitos básicos de redes neurais e HMM, detalhando as redes para reconhecimento de sinais dependentes do tempo, como as redes Elman e T-CombNET. No Capítulo 3 é descrito o sistema proposto, com os elementos para captura dos movimentos e identificação dos gestos. O Capítulo 4 apresenta os resultados obtidos na pesquisa. Finalizando o trabalho, no Capítulo 5 são apresentadas as conclusões e sugestões de trabalhos futuros.

Capítulo 2

Redes Neurais e HMM

O objetivo do trabalho é o estudo de redes neurais no reconhecimento dos gestos de maestro. Antes de descrever os detalhes do sistema proposto é apresentada uma breve introdução à redes neurais. Os resultados deste trabalho foram comparados com os obtidos com a utilização de HMM (*Hidden Markov Models*), e por este motivo também é apresentada uma breve introdução sobre HMM.

2.1 Fundamentos de redes neurais

As Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência.

Baseado no funcionamento de cada neurônio foi desenvolvido um modelo matemático. A combinação de vários destes neurônios artificiais compõe uma rede neural artificial, capaz de aprender. Uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento; já o cérebro de um mamífero pode ter muitos bilhões de neurônios.

Um neurônio é composto basicamente por dendritos, corpo do neurônio e um axônio, vide Figura 2-1. Os dendritos têm como função receber os estímulos emitidos por outros neurônios. O corpo do neurônio, também conhecido como *somma*, é

responsável por coletar e combinar as informações vindas dos dendritos. O axônio é responsável por transmitir os estímulos a outros neurônios.

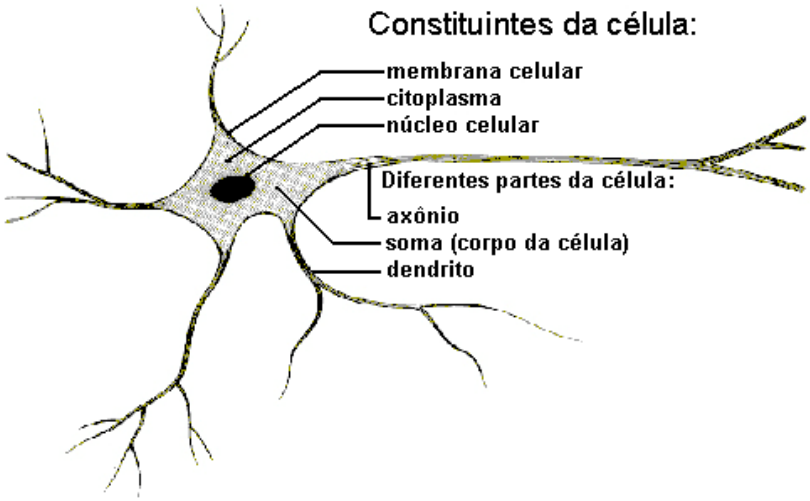


Figura 2-1: Esquema dos constituintes de uma célula neural.

Em 1943, McCulloch e Pitts fizeram os primeiros estudos com redes neurais. O modelo matemático escolhido para um neurônio artificial tenta reproduzir os elementos de uma célula neural. A Figura 2-2 faz uma analogia entra uma célula neural e o modelo matemático proposto McCulloch-Pits [15].

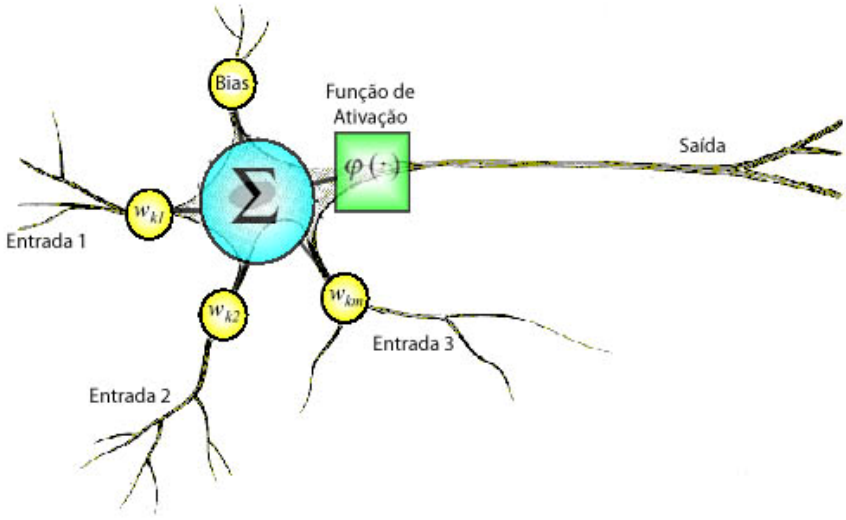


Figura 2-2: Analogia entre neurônio e seu modelo matemático.

Nos neurônios, a comunicação é realizada através de impulsos. Quando um impulso é recebido, o neurônio o processa e, passado um limite de ativação, dispara um segundo impulso que produz uma substância neurotransmissora, a qual flui do corpo celular para o axônio (que por sua vez pode ou não estar conectado a um dendrito de outra célula). O neurônio que transmite o pulso pode controlar a freqüência de pulsos aumentando ou diminuindo a polaridade na membrana pós sináptica.

Da mesma forma, um neurônio artificial recebe valores em suas diversas conexões de entrada, tendo cada conexão um peso próprio. As entradas são somadas e aplica-se uma função chamada “função de ativação” para limitar o valor da saída.

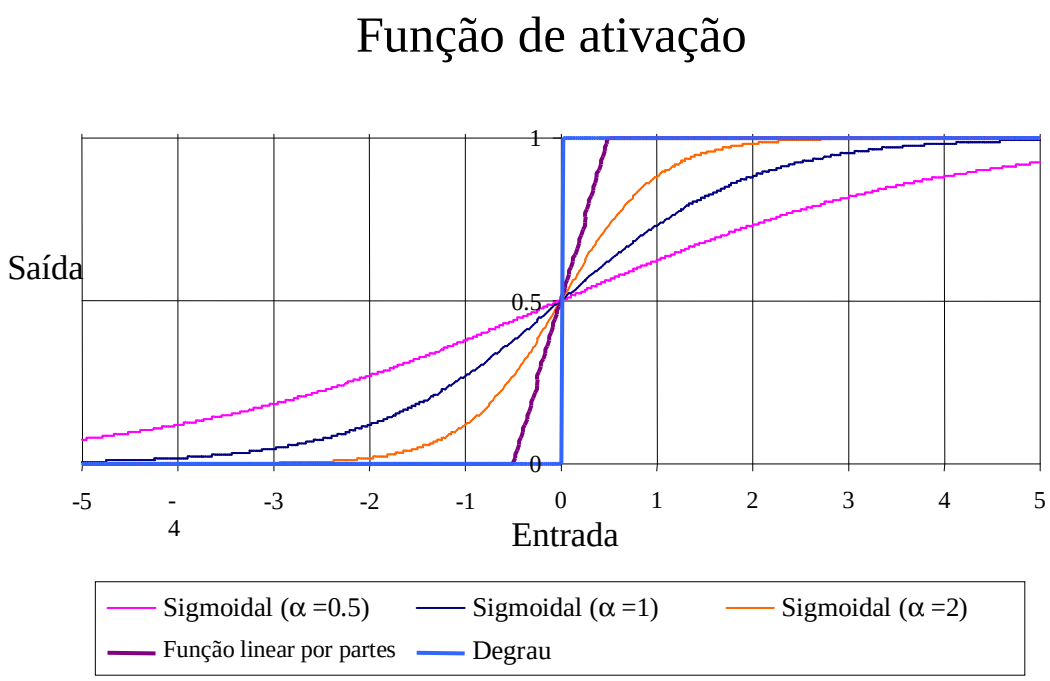


Figura 2-3: Funções de ativação.

A função de ativação pode ser de diversos tipos, alguns deles mostrados na Figura 2-3. O modelo de neurônio de McCulloch-Pits utiliza a função de ativação degrau , que é a função sigmoidal com $\alpha=\infty$. A função sigmoidal segue a equação 2.1.

$$\varphi(x) = \frac{1}{1+e^{-\alpha.x}} \tag{2-1}$$

O modelo do neurônio pode ser representado pela equação 2.2.

$$y_k = \varphi\left(\sum_{i=1}^{m} w_{ki} x_i + b_k\right) \tag{2-2}$$

Onde y_k é a saída do k-ésimo neurônio, w_{ki} o peso entre o neurônio k e a entrada x_i, b a polarização do neurônio e φ é a função de ativação.

Para facilitar a visualização das estruturas de redes neurais, a representação do neurônio é feita de forma simplificada como mostrado na Figura 2-4.

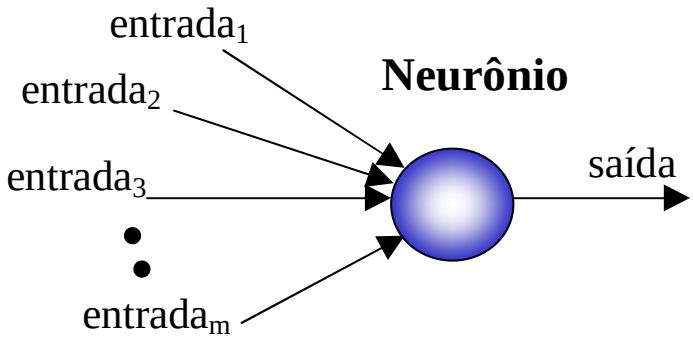


Figura 2-4: Representação de um neurônio

A conexão de vários neurônios forma uma rede neural. Uma rede neural artificial tem a capacidade de generalizar um problema, ou seja, é capaz de responder corretamente a uma entrada nunca vista antes por similaridade aos padrões já apresentados. Por isso é muito usada em aplicações de reconhecimento de padrões, classificação de dados e previsão.

2.2 Estruturas de redes neurais

Um dos objetivos da pesquisa sobre redes neurais na computação é desenvolver morfologias neurais matemáticas, não necessariamente baseadas na biologia, que podem realizar o processamento de alguns dados de forma desejada. As diferentes formas de conexões dos neurônios fazem a diferenciação entre os tipos de redes neurais e suas diferentes aplicações.

2.2.1 *Redes diretas (Feedforward)*

As redes diretas são representadas em camadas, sendo formadas por uma camada de neurônios de entrada, uma camada de saída e uma ou mais camadas intermediárias, conhecidas como camadas escondidas. A presença ou não de camadas escondidas, bem como o número de camadas escondidas é definido dependendo da complexidade e tipo do problema. Apenas uma camada intermediária é suficiente para aproximar qualquer função contínua e são necessárias no máximo duas camadas intermediárias, com um número suficiente de unidades por camada, para se produzir qualquer mapeamento.

Um exemplo de rede direta é a MLP (*Multi-Layer Perceptron*), cuja estrutura é mostrada na Figura 2-5.

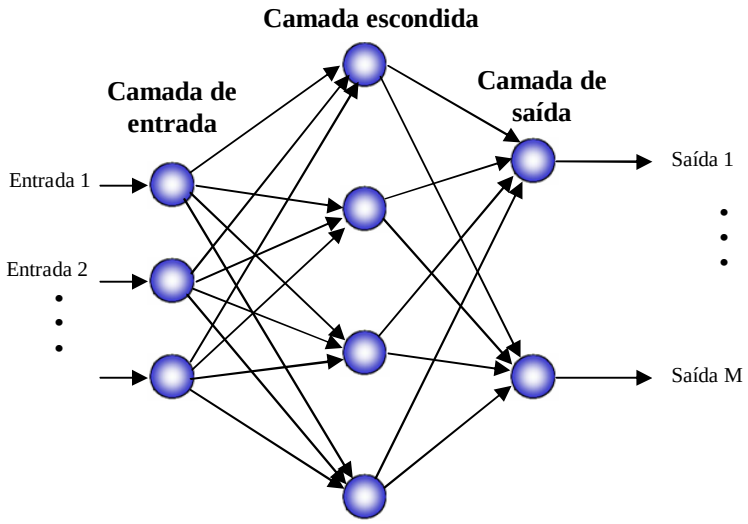


Figura 2-5: Rede neural direta MLP.

2.2.2 *LVQ (Learning Vector Quantization)*

Uma importante estrutura de rede neural é a LVQ proposta por Kohonen em 1990 [11]. O objetivo da rede LVQ é através da quantização dos vetores de entrada classificá-los em um determinado número de classes baseado em suas similaridades. A LVQ é rede composta de duas camadas com N neurônios de entrada e M neurônios de saída, conforme mostrado na Figura 2-6. Cada um dos N neurônios de entrada se conecta aos M neurônios de saída por conexões para frente. O número de classes em que a rede LVQ estará selecionando corresponde ao número de neurônios de saída.

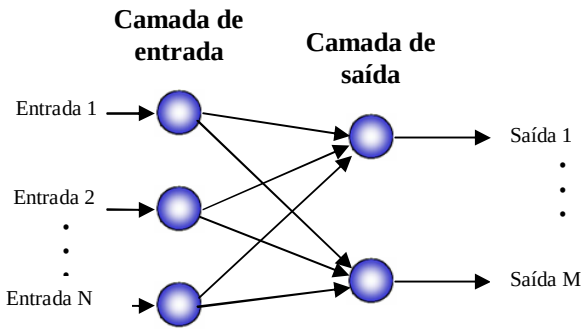


Figura 2-6: Estrutura da rede LVQ.

Os neurônios de saída da rede LVQ recebem idêntica informação na entrada, mas competem entre si para ser o único a ficar ativo. Cada neurônio se especializa numa área diferente do espaço de entrada e suas saídas podem ser usadas para representar a estrutura do espaço de entradas.

A função de ativação de cada neurônio é baseada na distância Euclidiana, conforme equação 2.3,

$$Y_j = \sum_{i=1}^{N} \left(X_i - W_{ij} \right)^2 \quad \textbf{(2-3)}$$

Onde Y_j é a saída do neurônio j, X_i é o vetor de entrada, W_{ij} é o peso entre o neurônio de entrada i e o neurônio de saída j.

O aprendizado pode ser feito de forma supervisionada ou não-supervisionada [33]. Nesta pesquisa foi utilizado o aprendizado não supervisionado, derivado do SOM (*Self Organizing Map*) [11].

2.3 Redes Neurais Temporais

As Redes neurais como a MLP e LVQ são estruturas capazes de processar de forma eficiente sinais não dependentes do tempo. No entanto, essas estruturas não são adequadas para processar sequências de eventos temporais. Surgem estruturas neurais que permitam processar tais sinais temporais.

2.3.1 *Redes não recorrentes*

Uma das soluções para resolver problemas temporais é a utilização de redes não recorrentes. Apesar de redes neurais como MLP e LVQ não terem bons resultados quando o problema é dependente do tempo, é possível adaptá-las para solucionar o problema. A solução mais comum é fixar uma janela no tempo e excitar a rede com os

valores desta janela, conforme a Figura 2-7. O problema desta solução é que a janela no tempo é fixa, e seqüências com duração maior que a janela não são reconhecidas.

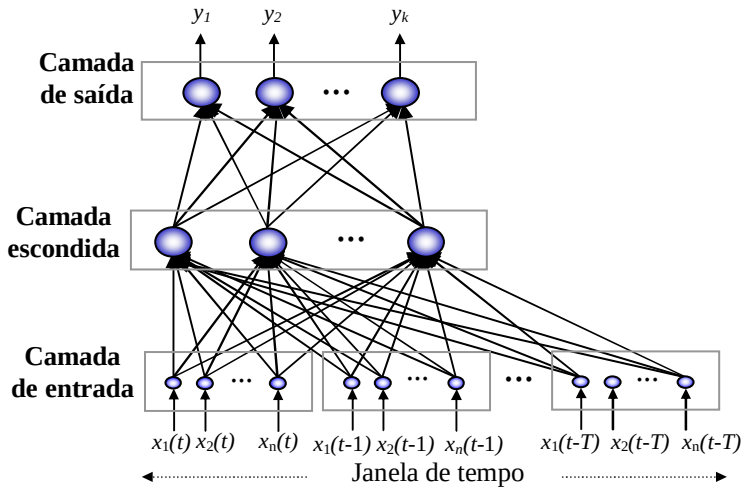


Figura 2-7: MLP com janela temporal.

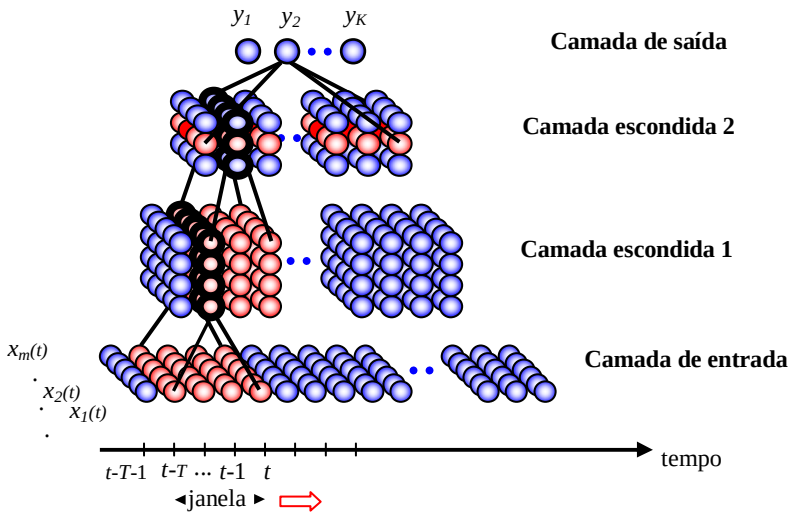


Figura 2-8: Rede neural com atraso no tempo.

Para solucionar o problema do tamanho limitado da janela, a estrutura TDNN (*Time Delay Neural Network*) [34] introduz uma segunda camada escondida, que recebe como entrada uma janela da primeira, assim como a camada de saída recebe os valores de uma janela da segunda camada escondida, como visto na Figura 2-8.

2.3.2 *Redes localmente recorrentes*

Nas redes neurais localmente recorrentes a estrutura do neurônio é modificada para que a rede consiga responder corretamente a estímulos temporais [35]. Na Figura 2-9, pode-se observar estágios de realimentação dentro da estrutura do neurônio, estes estágios funcionam como uma memória, armazenando o estado anterior do neurônio e utilizando-o em um próximo sinal de entrada.

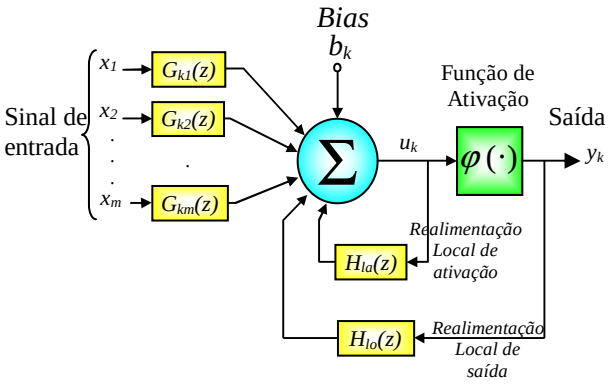


Figura 2-9: Neurônio recorrente.

Nessa estrutura de rede o processamento temporal é executado pelo neurônio. Geralmente são utilizadas redes neurais diretas, porém o modelo é genérico e pode ser utilizado em qualquer estrutura de rede.

2.3.3 *Redes totalmente recorrentes*

As Redes totalmente recorrentes, também conhecidas como redes simétricas, caracterizam-se por ter as conexões entre os neurônios nos dois sentidos. Assim como as redes diretas elas podem ter camadas de entrada, saída e camadas escondidas. Um exemplo da topologia de uma rede simétrica pode ser visto na Figura 2-10.

O modelo de neurônio utilizado é o de McCulloch-Pitts usando função sigmoidal de ativação. Tank e Hopfield [16] iniciaram os estudos com redes totalmente recorrentes e Williams e Zipser [17] propuseram um novo tipo de aprendizado para a

rede, porém ainda é difícil estabelecer critérios práticos para garantir a estabilidade da rede.

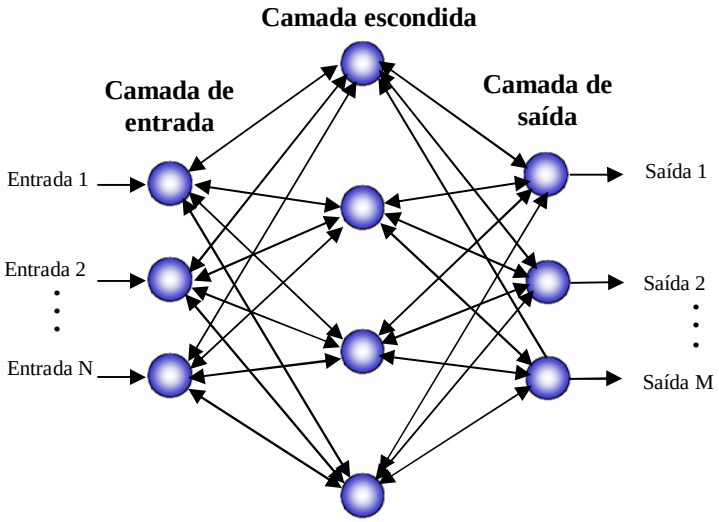


Figura 2-10: Topologia de uma rede simétrica.

2.3.4 *Redes parcialmente recorrentes*

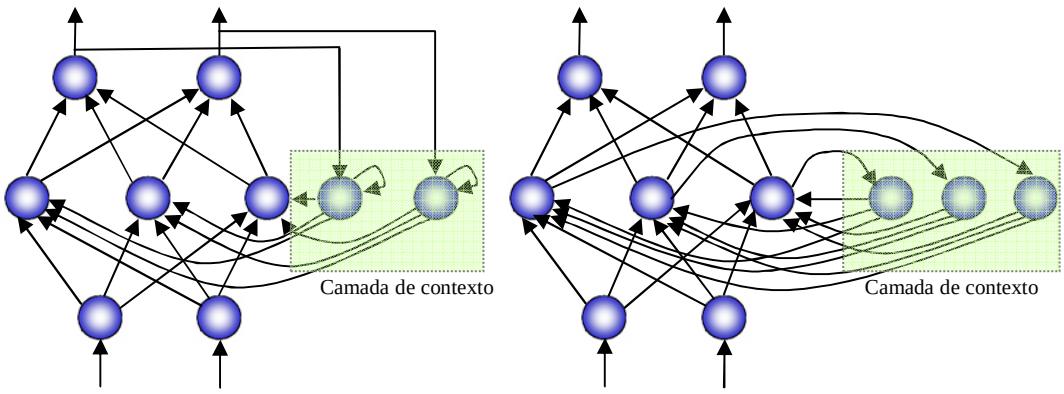
Tentando resolver os problemas de instabilidade e complexidade de treinamento das redes totalmente recorrentes foram feitos estudos utilizando *links* recursivos apenas em parte da rede. A solução proposta por Jordan e por Elman [40] foi criar uma nova camada chamada de camada de contexto, sendo tal camada responsável por guardar a informação temporal. A camada de contexto é uma das entradas da camada escondida.

Na solução proposta por Jordan a camada de contexto guarda informações da saída do instante anterior. Para o treinamento da rede Jordan é necessário a utilização de algoritmos específicos como o *backpropagation throught time.*

Na rede de Elman a camada de contexto guarda a informação dos neurônios da camada escondida. Os neurônios da camada de contexto têm função de ativação linear, ou seja, funcionam como memória. O peso entre a camada escondida e a camada de

contexto é fixado em 1, possibilitando a utilização do algoritmo comum de *backpropagation*.

As estruturas das redes de Jordan e Elman estão mostradas na Figura 2-11(a) e Figura 2-11(b), respectivamente.



(a) Rede Jordan **(b) Rede Elman**

Figura 2-11: Redes parcialmente recorrentes.

Esse tipo de rede se diferencia das demais soluções temporais por conseguir com que dados de todas as amostras anteriores influenciem no instante presente. Isso é possível devido à realimentação feita na camada escondida, que faz com que todas as amostras anteriores influenciem. Porém, quanto mais recente a amostra maior a influência no resultado. Por detectar o histórico temporal do sinal de forma mais completa que os outros tipos de redes, a estrutura parcialmente recorrente foi escolhida para o uso neste trabalho.

Em 1995, Kremer [9] demonstrou que apesar da simplicidade estrutural da rede de Elman, ela é suficientemente capaz de modelar informações temporais utilizando o modelo de neurônio de MacCuloch-Pitts e o algoritmo de *backpropagation*. Este fato motivou a utilização da rede de Elman neste trabalho.

2.4 Redes Neurais Compostas

Uma rede neural pode ser formada por uma combinação das estruturas clássicas anteriormente apresentadas. Dois exemplos de redes compostas são a CombNET-II e a T-CombNET.

2.4.1 *CombNET-II*

A rede CombNET-II é uma rede com estrutura em pente, formada por uma rede Tronco (*steam network*) e várias redes Galhos (*branch networks*), conforme demonstrado na Figura 2-12.

A rede tronco divide o espaço de entrada através de uma rede de Quantização Vetorial (VQ) em vários sub-espaços. A rede Tronco faz uma pré-seleção, escolhendo por qual rede Galho o sinal deve ser tratado. Cada neurônio de saída da rede Tronco é associado a uma rede Galho, o neurônio com a maior saída indica qual rede galho deve ser utilizada. Este estágio de pré-seleção faz com que cada uma das redes galho trabalhe com um número de classes reduzido.

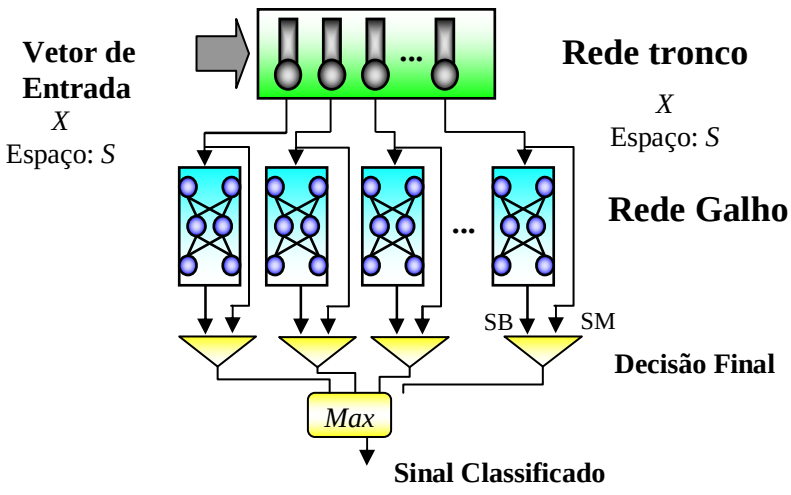


Figura 2-12: Estrutura da CombNET-II.

As redes Galhos consistem de redes Perceptron Multicamadas (MLP) de 3 camadas, e faz uma classificação refinada do vetor de entrada.

Os resultados obtidos na rede Tronco e Galho são processados num estágio de decisão, onde a saída é a classe em que o sinal de entrada pertence.

A principal vantagem da estrutura CombNET é a simplificação do estágio de treinamento. Uma vez que o espaço de entrada é pré-classificado, um grande problema é dividido em problemas menores, reduzindo o número de mínimos locais da superficie de erro nas redes Galhos, facilitando o treinamento pelo algoritmo de *Backpropagation* [12]. Esta filosofia permite que um problema com grande número de classes seja resolvido facilmente, reduzindo o tempo de treinamento e permitindo que uma boa solução seja encontrada, aumentando a taxa de reconhecimento.

2.4.2 *T-CombNET*

A rede CombNET-II utiliza estruturas não recorrentes e por este motivo não tem bons resultados em aplicações que envolvem sinais dependentes do tempo. Para que a CombNET-II seja utilizada para sinais dependentes do tempo é necessário que algumas alterações sejam feitas. A T-CombNET (*Temporal CombNET*) é uma variante da CombNET-II para sinais temporais.

Assim como a rede CombNET-II, a T-CombNET, mostrada na Figura 2-13, é composta por um estágio pré-classificatório, capaz de separar o espaço de entrada em sub-espaços, e por várias redes chamadas de redes Galhos. Os resultados obtidos pelas duas redes são analisados para a decisão final.

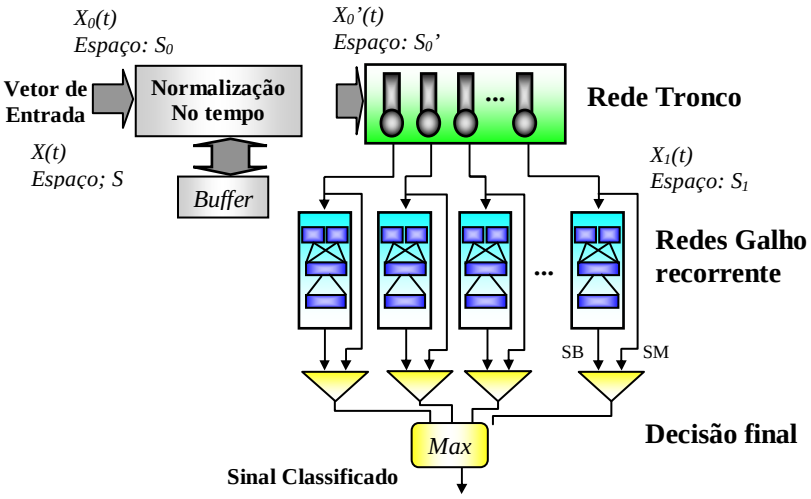


Figura 2-13: Estrutura da rede T-CombNET.

O processo pré-classificatório é não supervisionado e, portanto, o número de redes Galhos presente no sistema não é pré-definido.

A estrutura T-CombNET necessita de um estágio de normalização no tempo. Esta normalização no tempo é feita no sinal que será aplicado à rede tronco. A rede utilizada como tronco é a LVQ [11], que necessita de um vetor de entrada com dimensão fixa, e a rede galho é necessariamente uma rede temporal, como por exemplo a Elman [9], utilizada neste trabalho.

Os diversos estudos em redes neurais geraram inúmeras estruturas de redes, com diferentes aplicações. Neste trabalho as redes neurais utilizadas são as redes Elman e T-CombNET. A escolha das redes se fundamentou na necessidade da identificação temporal e nos bons resultados apresentados em pesquisas anteriores no reconhecimento de gestos.

2.5 Treinamento de rede neural

As redes neurais podem ser treinadas de forma supervisionada ou não supervisionada. Os dois métodos são descritos a seguir.

2.5.1 *Treinamento supervisionado*

A maioria das estruturas de redes neurais são treinadas de forma supervisionada. No treinamento supervisionado a rede recebe os vetores de entrada e os resultados desejados na saída da rede.

A tarefa do treinamento da rede é ajustar os pesos entre os neurônios de forma que a saída seja a desejada. Uma rede bem treinada é capaz de generalizar um problema, ou seja, não é necessário que a rede conheça todas as possibilidades de entrada para obter a saída. Nosso cérebro funciona de forma similar em muitos casos. Por exemplo, você consegue identificar que determinado ser vivo é uma árvore, mesmo nunca tendo visto aquele espécime em particular, uma vez que você já foi apresentado a outros tipos de árvores no passado. Este é o processo de generalização do aprendizado.

Para conseguir treinar a rede de forma a generalizar um problema é necessário escolher de forma cuidadosa o banco de dados de treinamento, ele deve abranger as diversas variações com algumas repetições. Um banco de dados pequeno pode não ser suficiente para o treinamento da rede e um banco de dados muito grande pode elevar demais o custo computacional para o de treinamento da rede.

Existem vários algoritmos para realizar o treinamento da rede e o mais conhecido deles é o de *backpropagation.*

Backpropagation

Durante o treinamento com o algoritmo *backpropagation*, a rede opera em uma sequência de dois passos. Primeiro, um padrão é apresentado à camada de entrada da rede. A atividade resultante flui através da rede, camada por camada, até que a resposta seja produzida pela camada de saída. No segundo passo, a saída obtida é comparada à saída desejada para esse padrão particular. Se esta não estiver correta, o erro é calculado. O erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo modificados conforme o erro é retropropagado.

Este processo é repetido para cada um dos padrões selecionados para treinamento e este ciclo completo é denominado de época. Várias épocas são necessárias para que a rede chegue aos resultados desejados. O treinamento é finalizado quando o erro de saída atinge valores abaixo de um valor estipulado ou quando atinge-se um pré-determinado número de épocas.

2.5.2 *Treinamento não-supervisionado*

No treinamento não-supervisionado somente os padrões de entrada estão disponíveis para o treinamento da rede. O treinamento não supervisionado é utilizado nas redes LVQ.

O treinamento é feito por competição. O padrão é apresentado à camada de entrada da rede. A função de ativação calcula a distância Euclidiana entre o vetor de entrada e os pesos sinápticos. O neurônio que tiver a menor distância Euclidiana vence e tem seus pesos atualizados de forma a se aproximar mais do padrão de entrada.

Desta forma consegue-se treinar a rede LVQ para separar os padrões de entradas em M grupos.

2.6 HMM (*Hidden Markov Models*)

O HMM é um modelo estatístico baseado nas Cadeias de Markov, estrutura amplamente utilizada para a modelagem de processos estocásticos. As Cadeias de Markov tem sua fundamentação teórica baseada no estudo de vetores de probabilidades e algumas propriedades da Álgebra Linear (Autovetores e Autovalores). Além destas características, a principal idéia quando se trata de processos Markovianos é assumir que, dada uma seqüência de eventos existe dependência entre alguns destes. Estas dependências podem ser classificadas como de:

· 1ª ordem: A probabilidade de ocorrência do *k*-ésimo evento é depende do evento imediatamente anterior.

· 2ª ordem: A probabilidade de ocorrência do *k*-ésimo evento é depende dos dois eventos imediatamente anteriores.

· *n*-ésima ordem: A probabilidade de ocorrência do *k*-ésimo evento é depende dos *n* eventos imediatamente anteriores.

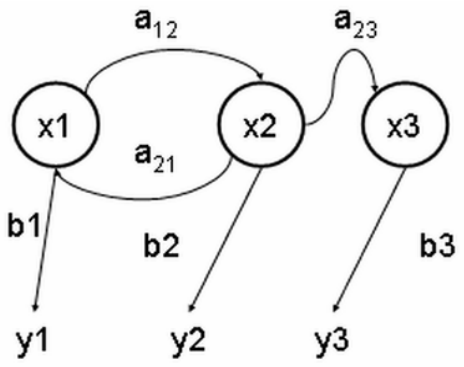


Figura 2-14: Estados de transição.

Dessa forma pode-se montar os estados de transição de um problema, como descrito na Figura 2-14. A probabilidade da transição de um estado *i* para um estado *j* é a_{ij}, os estados são representados por *x*, as saídas observáveis por *y* e *b* é a probabilidade de ocorrer determinada saída.

Uma vez definida a matriz de probabilidades é possível identificar a ocorrência de um padrão conhecido. Por este motivo HMM é bastante usado para reconhecimento de voz e também pode ser usado em reconhecimento de gestos [36].

O tempo de treinamento de HMM é mais curto que redes neurais. A desvantagem de HMM é que é necessário ter o sinal completo para que seja possível a identificação do gesto.

Existem três problemas básicos que devem ser resolvidos para que modelo possa ser utilizado em aplicações do mundo real.

Esses problemas são os seguintes:

Problema 1 (problema de avaliação): Dado a seqüência de observação $O = (o_1, o_2, ..., o_T)$, e o modelo $\lambda = (A, B, \pi)$, como calcular eficientemente $P(O \mid \lambda)$ a probabilidade da seqüência de observações, dado o modelo?

A maneira mais direta de calcular a probabilidade de uma seqüência de observações é através da enumeração de todas as possíveis seqüências de estados de tamanho T (o número de observações), pela seguinte expressão:

$$P(O \mid \lambda) = \sum_{q_1, q_2 \ldots q_T}^{\pi_{q_1} b_{q_2}(o_1) a_{q_1 q_2}} b_{q2}(o_2) \ldots a_{q_{r-1} q_T} b_{qT}(o_T) \tag{2-4}$$

Algoritmos como o de *backward* e *forward* também são usados para resolver o problema, e são de menor custo computacional.

Problema 2 (problema da busca da melhor seqüência de estados): Dado a seqüência de observações) $O = (o_1, o_2, ..., o_T)$, e o modelo λ, como escolher uma seqüência de estados correspondente $Q = (q_1, q_2, ..., q_T)$?

Este problema geralmente é resolvido usando um procedimento próximo ao ótimo, o algoritmo de Viterbi [37][38].

Problema 3 (problema de treinamento): Como ajustar os parâmetros do modelo $\lambda = (A, B, \pi)$ para maximizar $P(O \mid \lambda)$?

Não existe uma maneira conhecida de resolver analiticamente o conjunto de parâmetros do modelo que maximiza a probabilidade da seqüência de observações de uma maneira fechada. Entretanto, pode-se escolher $\lambda = (A, B, \pi)$ tal que sua probabilidade $P(O \mid \lambda)$ é localmente maximizada usando um procedimento iterativo tal como o método de Baum-Welch [39].

Capítulo 3

Sistema proposto

O objetivo do trabalho foi reproduzir computacionalmente a identificação feita por um músico dos gestos de um maestro. No sistema proposto a percepção do olho humano é substituída por uma câmera e o processamento do cérebro por um computador, como mostrado na Figura 3-1.

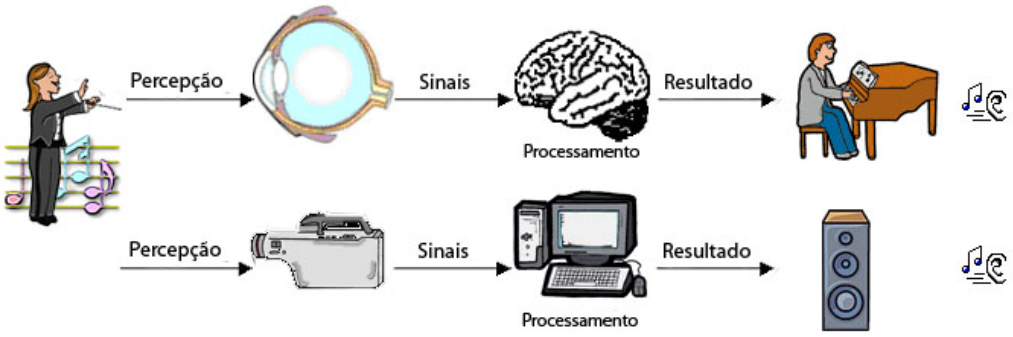


Figura 3-1: Comparação do sistema real e o computacional proposto

O músico recebe como entrada as notas de uma partitura e os gestos do maestro. As notas da partitura indicam a melodia da música e os gestos do maestro gerenciam a forma com que essas notas são tocadas, de forma a uniformizar a orquestra. Com a partitura e um maestro, um músico tem todas as informações necessárias para tocar a música, mesmo sem nunca a ter ouvido.

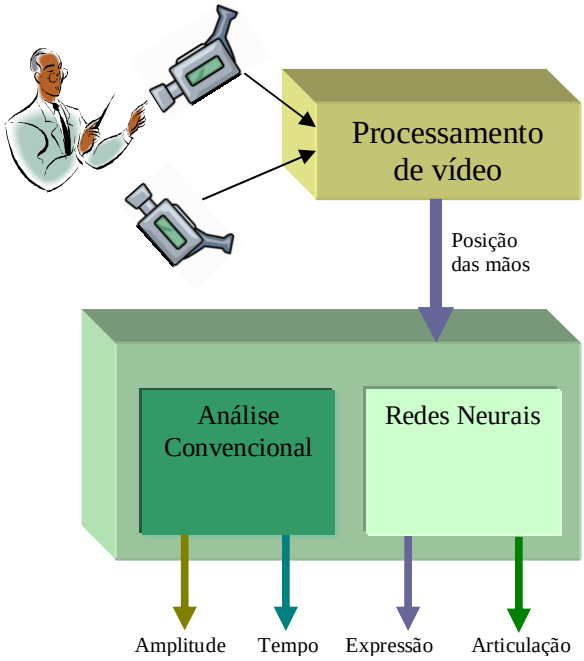


Figura 3-2: Sistema proposto

O sistema proposto, conforme Figura 3-2, pode ser dividido em três partes principais: a aquisição e processamento de vídeo; análise convencional para a extração de dinâmica e tempo; e o reconhecimento dos gestos utilizando redes neurais. As três partes são detalhadas nas próximas seções.

3.1 Aquisição e processamento de vídeo

A aquisição dos vídeos foi feita utilizando duas câmeras de vídeo digitais de baixo custo, captando imagens a uma freqüência de 25 quadros por segundo e com 352x288 pixels de resolução. As duas câmeras foram posicionadas com um ângulo de 90 graus entre si, captando os gestos do maestro na posição frontal e lateral como mostra a Figura 3-3. Os vídeos foram capturados na Universidade McGill, Montreal,

Canadá, por Paul Kolesnik. O maestro que executou os gestos foi Lana Lysogor, doutora em regência.

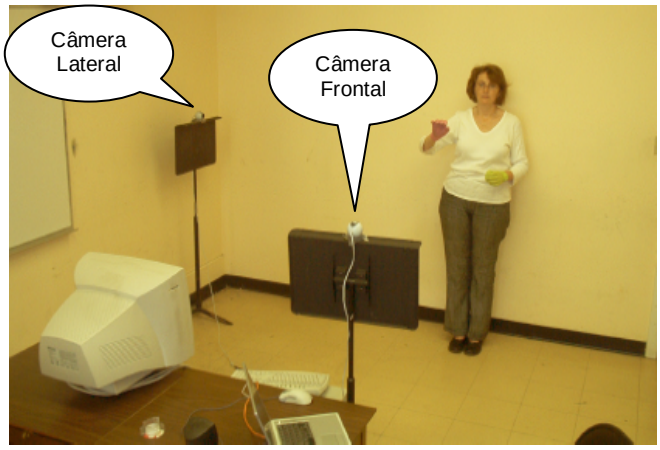


Figura 3-3: Posicionamento das câmeras na captura dos gestos.

Para facilitar a identificação da mão do maestro foi utilizada uma luva, de tom semelhante ao da pele, para uniformizar a cor da mão. Utilizando como base a cor da luva do maestro um sistema de *tracking* foi desenvolvido para a identificação da posição da mão. A Figura 3-4 mostra dois quadros com movimento *staccato* onde a trajetória foi capturada pelo sistema de *tracking*. A linha amarela mostra o primeiro pulso e a linha verde o segundo pulso.

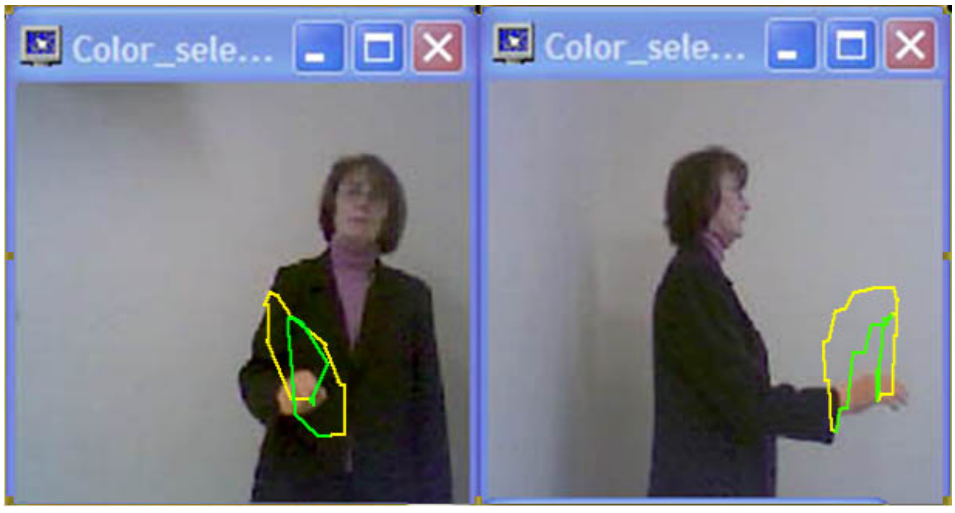


Figura 3-4: Movimento *staccato* de 2 pulsos.

A captação e o processamento dos vídeos para a extração da posição das mãos foram feitos no programa Eyesweb [14].

A identificação da posição da mão foi feita pela cor. São identificados quais *pixels* têm cor semelhante à desejada e então calculado o baricentro da região. A busca por *pixels* da cor desejada não foi feita em toda a imagem, para diminuir o processamento e para diminuir a chance de identificar *pixels* não desejados. Para tanto, a busca foi feita apenas em uma área delimitada por um quadrado com centro no ponto identificado como a luva no quadro anterior, como indicado na Figura 3-5.

Figura 3-5: Quadro de busca da posição da mão.

O mesmo processamento foi feito para os vídeos captados por ambas as câmeras gerando quatro coordenadas, sendo uma delas redundante.

Após a identificação da posição das mãos em cada quadro, as coordenadas 3D foram armazenadas em arquivo texto, para que o treinamento das redes fosse efetuado.

3.2 Medida de *tempo* e *dinâmica*

A extração do *tempo* e *dinâmica* foi baseada na posição de um pulso no tempo e a posição da mão no instante de cada pulso. O instante em que ocorre o pulso é definido quando a mão direita do maestro muda o movimento de descida para subida. A Figura 3-6 mostra a identificação dos pulsos em uma seqüência de quadros.

Com a posição das mãos em cada quadro foi implementado um programa para a identificação das transições de velocidade vertical negativa para positiva.

A velocidade em que os pulsos ocorrem define o tempo, calculado em batidas por minuto (bpm).

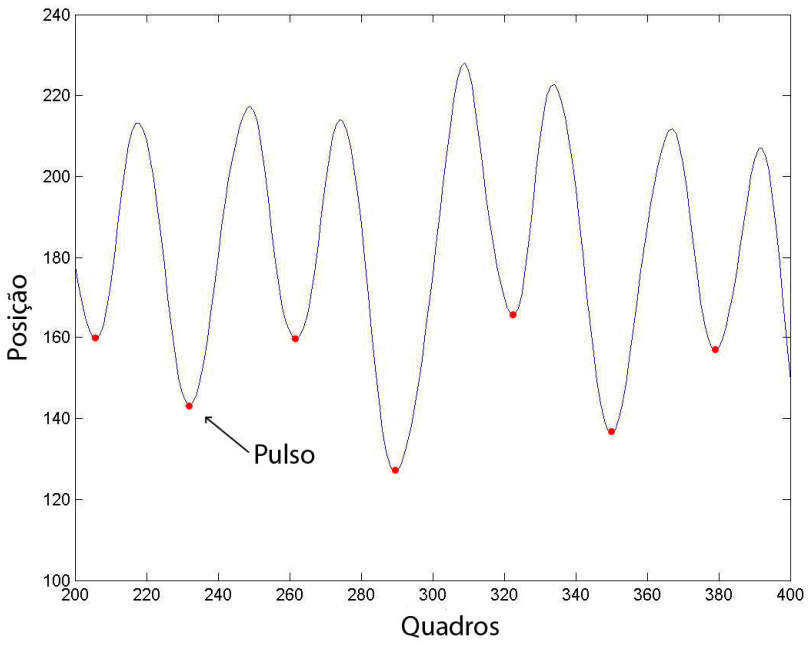


Figura 3-6: Identificação dos pulsos.

A dinâmica é baseada na amplitude dos movimentos do maestro. A amplitude foi calculada pela diferença entre os pontos de máximo e mínimo da coordenada vertical entre dois pulsos consecutivos.

Como a amplitude para cada pulso dentro do compasso é naturalmente diferente, o músico necessita que a variação da amplitude se repita por mais de um pulso para que consiga identificar a variação. Por este motivo, e para amortizar indesejáveis e pequenas variações na execução do movimento pelo maestro, a variação de amplitude considerada foi uma média da amplitude instantânea durante 100 amostras, ou seja, 4 segundos.

O tempo de 4 segundos foi usado por ser intervalo suficiente e mínimo para abranger um compasso completo do movimento.

O mesmo critério de média foi utilizado também para o tempo, minimizando as pequenas e indesejáveis variações da posição do pulso no tempo durante o compasso.

3.3 Classificação de gestos

O reconhecimento de gestos pode ser feita utilizando diversas técnicas. As mais populares são os HMM (*Hidden Markov Models*) e as redes neurais. Este trabalho se propôs a analisar a utilização de redes neurais para a identificação dos gestos de maestro, fazendo uma comparação com resultados obtidos por HMM na pesquisa de Kolesnik [8].

Por tratar-se de uma identificação de uma variação no tempo, faz-se necessária a utilização de redes neurais temporais. Dentre as diversas estruturas apresentadas no Capítulo 2, foi decidido pelo uso das redes parcialmente recorrentes de Elman [9] e da T-CombNET [10]. Ambas as redes foram escolhidas pelos bons resultados apresentados em pesquisas anteriores no reconhecimento de gestos.

Capítulo 4

Resultados Experimentais

A pesquisa foi feita em duas etapas, numa primeira etapa foram feitos testes com um banco de dados simples, com gestos básicos, e no segundo os testes foram feitos utilizando o mesmo sistema, porém com um banco de dados real, com um maestro realizando gestos da gramática básica.

4.1 Experimento 1

O experimento 1 foi feito para desenvolvimento do sistema de reconhecimento, assim como aferição do código implementado. Para que este fim fosse atingido foi montado um banco de dados com vídeos captados nos mesmos padrões que seriam utilizados no experimento final. O sistema de reconhecimento usado foi o mesmo descrito no Capítulo 3.

4.1.1 *Banco de dados*

Para este experimento foram escolhidos apenas 3 movimentos, com características bem distintas entre si. A posição absoluta 2D da mão de cada gesto pode ser observada na Figura 4-1. Observe que os 3 movimentos têm trajetórias bem diferentes e de fácil identificação visual.

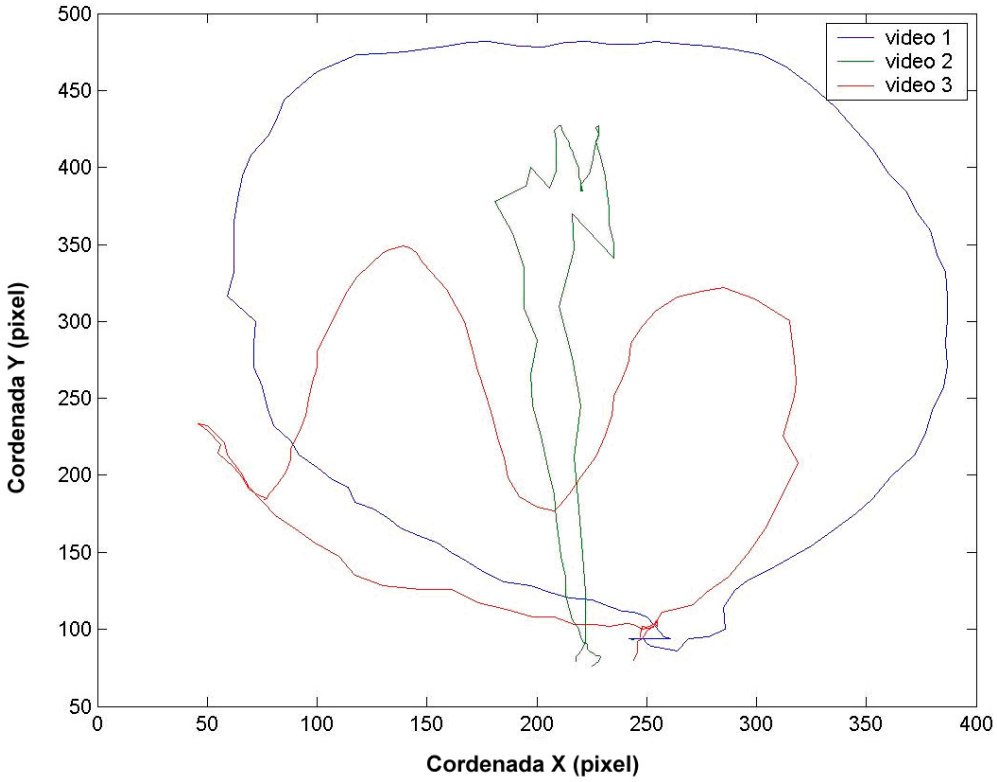


Figura 4-1: Trajetória 2D dos 3 movimentos

Os movimentos foram feitos utilizando uma luva, como mostrado na Figura 4-2, e foram repetidos 9 vezes cada um, resultando em um banco de dados de 27 vídeos.

Figura 4-2: Quadro de um vídeo capturado

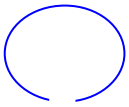
Dos 9 vídeos de cada gesto, 5 foram utilizados para treino da rede e 4 para teste da rede. Cada vídeo tinha em média 6 mb e foi capturado por uma câmera Logitech QuickCam Messenger USB.

4.1.2 *Resultados*

O objetivo deste experimento foi o desenvolvimento do sistema a ser utilizado no reconhecimento de gestos de maestro, apresentado no item 4.2 . Portanto, o produto final do experimento esperado não foi tanto o grupo de resultados obtido no reconhecimento dos gestos, facilitado pela diferença entre os gestos, mas sim a validação das etapas do sistema de reconhecimento proposto.

O sistema de reconhecimento resultante e aferido por este experimento está descrito no Capítulo 3. O resultado obtido no reconhecimento dos gestos foi de 100%, conforme apresentado na Tabela 4-1.

Tabela 4-1: Resultados do experimento 1.

Gesto	**Trajetória do movimento**	**Vídeos de treino**	**Vídeos de teste**	**Taxa de reconhecimento**
Gesto 1		5	4	100%
Gesto 2		5	4	100%
Gesto 3		5	4	100%
Total / Média		15	12	100%

Os testes foram feitos utilizando a rede Elman com 2 neurônios de entrada, 30 na camada escondida e 3 na de saída. Para o treinamento da rede utilizou-se a coordenada absoluta da mão sem nenhum processamento.

A grande diferença entre os movimentos dos gestos fez o experimento não suficiente para a avaliação do sistema, porém o resultado de reconhecimento total obtido no experimento valida a metodologia proposta e sinaliza que o sistema pode apresentar bons resultados para testes com gestos mais complexos e em maiores números, objetivo do experimento 2.

4.2 Experimento 2

O experimento 2 consistiu na utilização do sistema descrito no Capítulo 3 para a identificação de gestos de maestro.

Este experimento foi dividido em duas partes: a análise de tempo e dinâmica, e o reconhecimento dos gestos. Para a análise de tempo foi feita a identificação da posição dos pulsos no tempo, e a posição da mão em cada pulso foi utilizada na análise da dinâmica, conforme descrito no item 3.2 . Para o reconhecimento de gestos foram utilizadas redes neurais parcialmente recorrentes, conforme descrito no item 3.3 .

No reconhecimento de gestos foi feita uma comparação entre as duas redes propostas e o mesmo experimento realizado com a identificação sendo feita por HMM [8].

Os resultados foram divididos em: articulação, tempo, gestos da mão direita e gestos da mão esquerda. Um item final mostra resultados obtidos na segmentação dos gestos.

4.2.1 *Banco de dados*

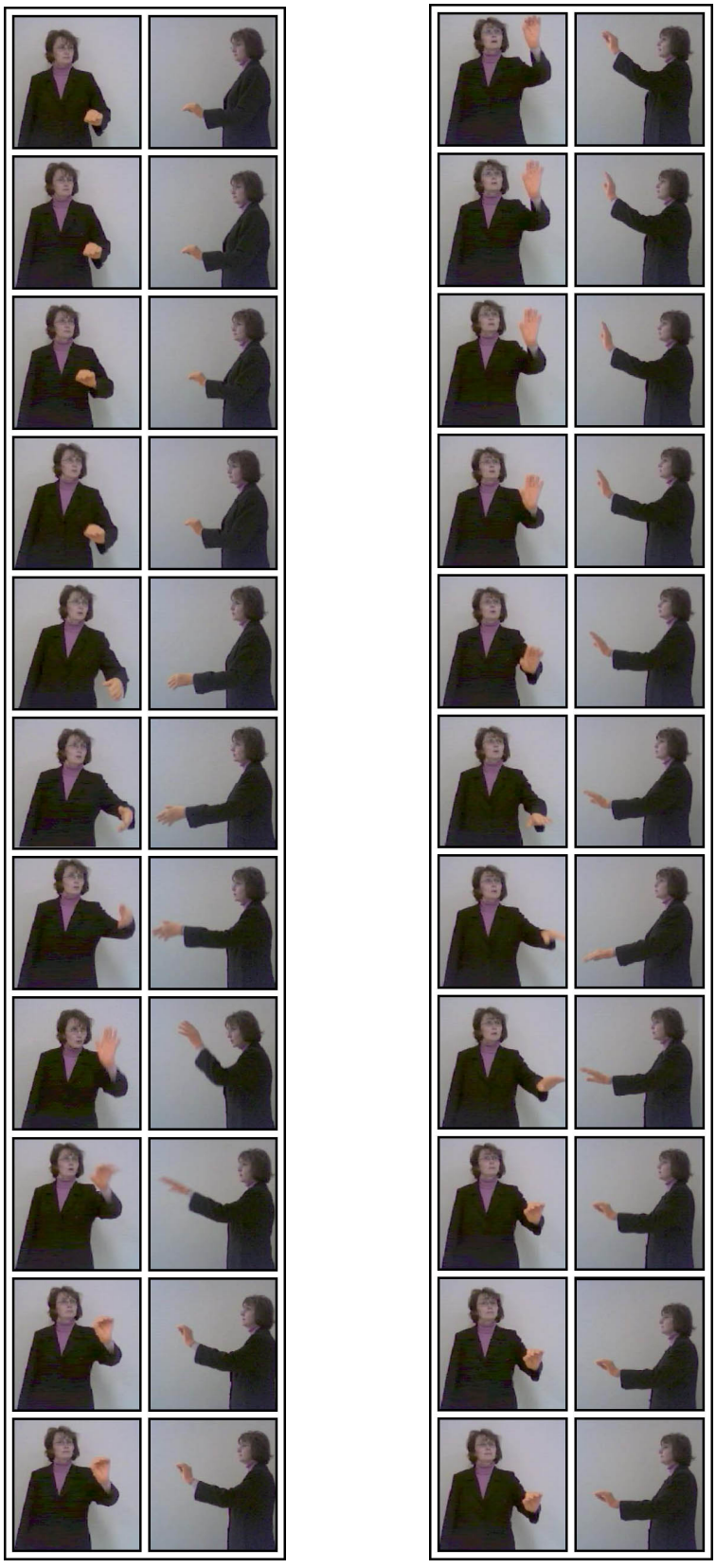
Os gestos do maestro expressam articulação, tempo, expressão e dinâmica. Em sua forma mais comum os gestos de expressão são executados pela mão esquerda e os de articulação, tempo e dinâmica pela mão direita, porém, em alguns casos específicos alguns gestos de expressão podem ser representados pela mão direita.

Para a pesquisa foram escolhidos gestos padronizados que representem todos estes movimentos. Os vídeos capturados para a pesquisa podem ser divididos em 3 grupos, como mostra a Tabela 4-2. O primeiro grupo foi composto por 5 gestos, executados pela mão esquerda do maestro, indicando expressão. No segundo grupo foram 3 gestos da mão direita indicando expressão da música. No terceiro grupo de gestos foram 7 vídeos da mão direita, indicando articulação, tempo e dinâmica.

Tabela 4-2: Banco de dados do experimento 2.

Tipo dos gestos	Mão que executa	Número de classes	Número de vídeos de treino	Número de vídeos de teste
Expressão	Esquerda	5	75	50
	Direita	3	60	30
Articulação, tempo e dinâmica	Direita	7	140	70

A Figura 4-3 mostra quadros extraídos de vídeos de expressão, executados pela mão esquerda. A Figura 4-4 mostra quadros extraídos de vídeos de articulação, executados pela mão direita. Cada vídeo tinha em média 2 mb e foi capturado por uma câmera Logitech QuickCam Messenger USB.



(a) (b)

Figura 4-3: (a) Quadros de um vídeo do gesto *crescendo*+corte (b) Quadros de um vídeo do gesto *diminuendo*+corte.

(a) (b)

Figura 4-4: (a) Quadros de um vídeo do gesto *legato* 2 pulsos/compasso (b) Quadros de um vídeo do gesto *marcato* 2 pulsos/compasso.

Após o *tracking* dos vídeos, cada gesto foi submetido a um pré-processamento e armazenado em um arquivo texto contendo as coordenadas da mão em cada quadro. Estes arquivos foram utilizados como base de dados para o treinamento e testes da rede.

O pré-processamento feito nos dados de posição da mão após o *tracking* gera o vetor de características. Foram feitos testes com 4 tipos de vetores de características: posição absoluta, posição absoluta filtrada, velocidade e direção.

Posição absoluta

Na posição absoluta não foi feito nenhum processamento, os dados capturados pelo *tracking* foram enviados diretamente para a rede neural.

A coordenada absoluta é dependente da posição relativa da câmera ao maestro, qualquer mudança na posição que o maestro executa o movimento pode resultar na não identificação do gesto.

Posição absoluta filtrada

Com a coordenada absoluta foi feito um filtro passa-baixas para eliminar as pequenas variações na execução dos gestos, e o resultado desta filtragem também foi utilizado no treinamento da rede.

O filtro utilizado foi um filtro de Butterworth de décima ordem e freqüência de corte normalizada de 0,2.

A filtragem passa-baixas na coordenada absoluta faz com que pequenas alterações no gesto do maestro sejam ignoradas na identificação. Porém, o problema da dependência da posição da câmera e do maestro presente na coordenada absoluta não é resolvido.

Velocidade

A informação de velocidade é calculada de acordo com

$$Vx_k = X_k - X_{(k-1)} \tag{4-1}$$

$$Vy_k = Y_k - Y_{(k-1)} \tag{4-2}$$

$$Vz_k = Z_k - Z_{(k-1)} \tag{4-3}$$

onde X_k , Y_k *e* Z_k, são as coordenadas do ponto no instante *k*, e X_{k-1}, Y_{k-1}, Z_{k-1} as coordenadas no instante anterior *k*-1.

A velocidade não é dependente da posição em que o movimento foi executado, facilitando a reprodução do movimento.

Direção

A informação de direção é calculada por

$$M_k = \sqrt{Vx_k^{\,2} + Vy_k^{\,2} + Vz_k^{\,2}} \tag{4-4}$$

$$Dx_k = \frac{Vx_k}{M_k} \tag{4-5}$$

$$Dy_k = \frac{Vy_k}{M_z} \tag{4-6}$$

$$Dz_k = \frac{Vz_k}{M_k} \tag{4-7}$$

onde Vx_k,Vy_k *e* Vz_k, são as componentes da velocidade no ponto no instante *k*, e M_k corresponde ao módulo do vetor velocidade. Definindo assim, *D* um vetor unitário no espaço 3-D que define apenas a direção do movimento.

A direção é independente da velocidade e, assim como a velocidade, não é dependente da posição em que o movimento foi executado.

No reconhecimento de gestos os testes com as coordenadas absolutas filtradas e velocidade obtiveram melhores resultados. Devido à dificuldade de reprodução de um

movimento utilizando a coordenada absoluta filtrada, os testes finais foram feitos utilizando os dados de velocidade como entrada da rede neural.

A comparação entre os tipos de dados a serem usados foi empírica. O fato da velocidade obter melhores resultados do que direção indica que a informação de velocidade, ignorada na direção, é importante na identificação do movimento.

Para a extração do tempo e dinâmica foi utilizada a coordenada absoluta filtrada.

4.2.2 *Tempo*

Para a análise do *tempo* foram utilizados 7 vídeos da mão direita, e empregada a análise convencional descrita no item 3.2 .

Para a extração do *tempo* é necessária a identificação de todos os pulsos presentes no vídeo. A identificação dos pulsos foi feita por um algoritmo que verifica a mudança da direção de velocidade de negativa para positiva no eixo vertical, conforme apresentado no item 3.2 . A taxa de reconhecimento de pulsos média foi de 100%.

Com a posição de todos os pulsos no tempo, foi calculado o *tempo* em cada instante e o *tempo* médio em um intervalo de 4 segundos. A Figura 4-5 apresenta os resultados do *tempo* instantâneo e médio para os 7 vídeos. A linha azul representa o tempo entre dois pulsos e a linha verde é a média entre 100 amostras (4 segundos).

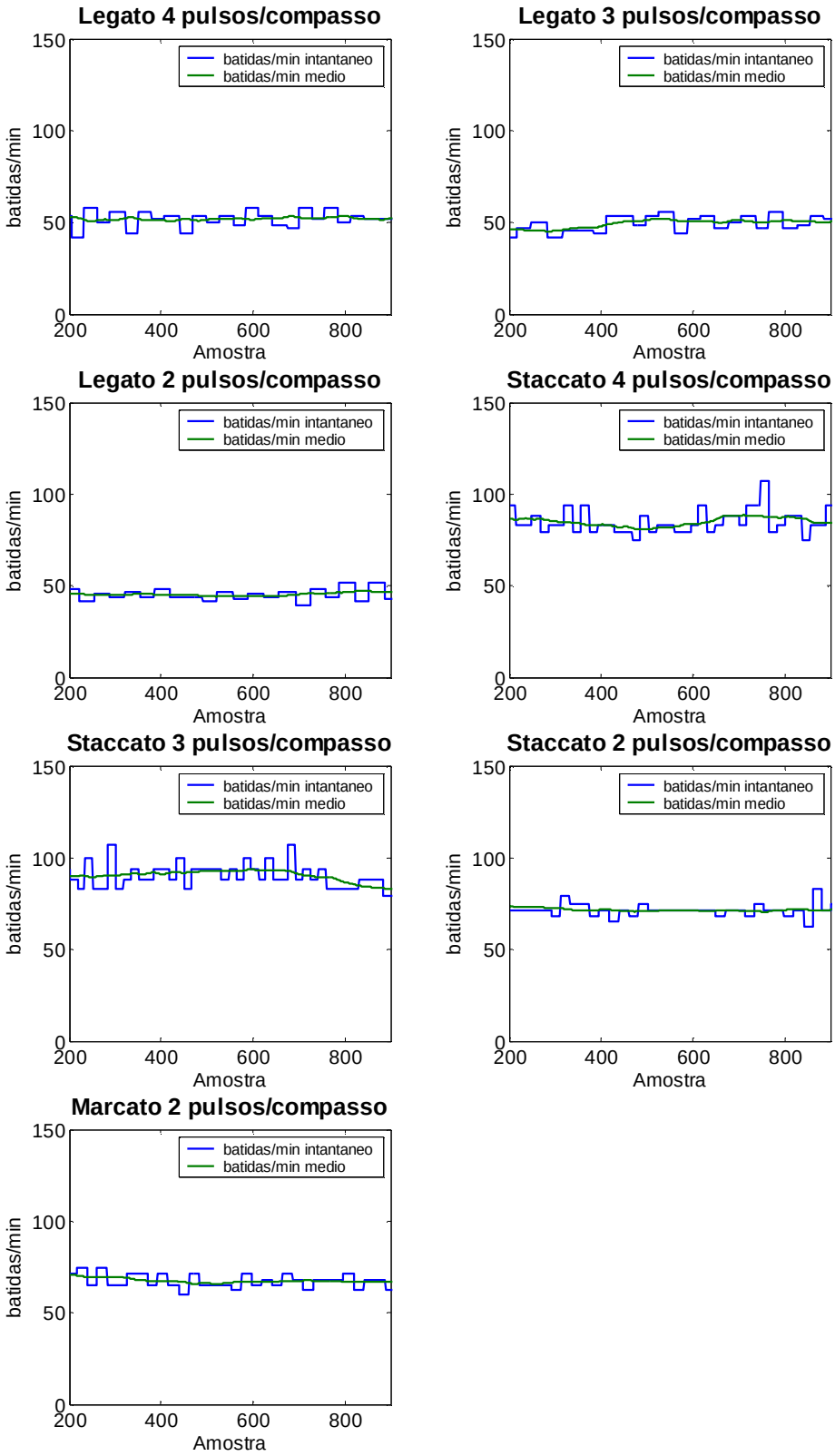


Figura 4-5: Tempo instantâneo e médio nos 7 vídeos.

A Figura 4-6 reúne os resultados do tempo médio para os 7 vídeos.

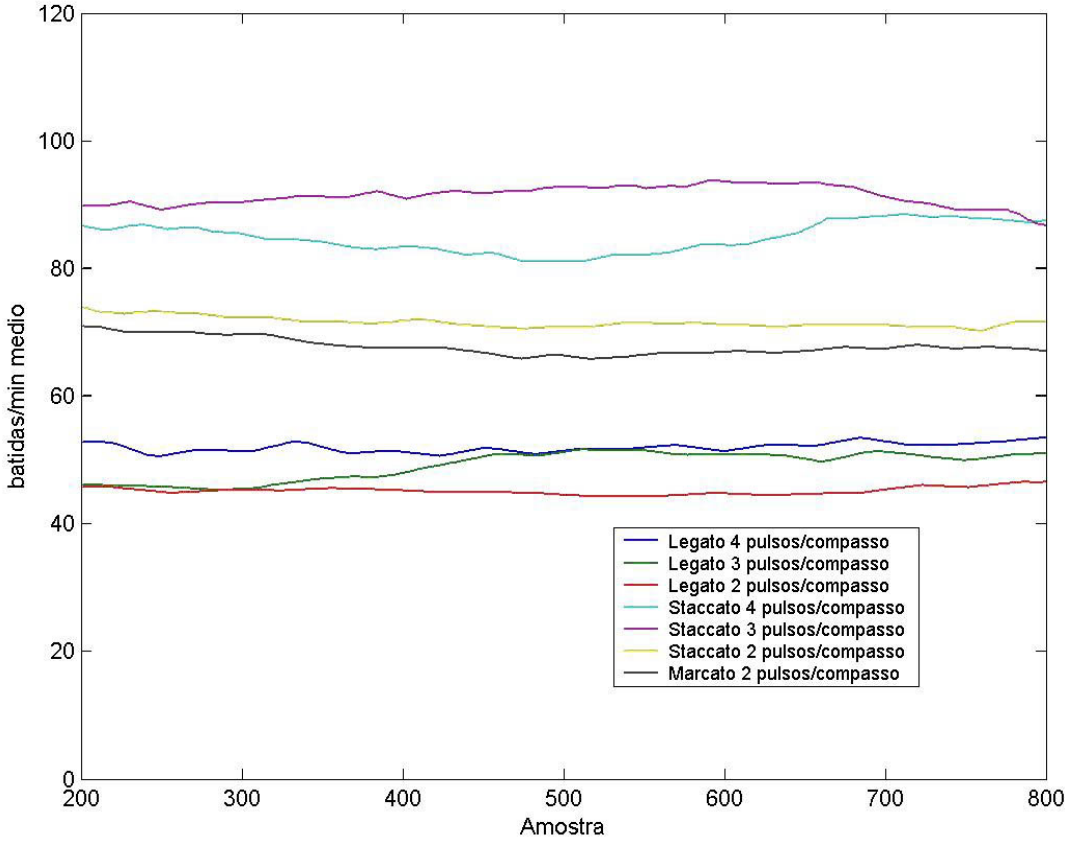


Figura 4-6: Batidas/min médio em trecho dos 7 vídeos.

Percebe-se que a variação no tempo em todos os vídeos foi muito pequena indicando uma intenção do maestro de manter o tempo durante a execução da música. Essa constância no *tempo* é confirmada ao observar os vídeos.

4.2.3 *Dinâmica*

Para a análise da dinâmica foram utilizados 7 vídeos da mão direita e empregada a análise convencional descrita no item 3.2 . Assim como para o *tempo,* é necessária a identificação dos pulsos, e a taxa de reconhecimento de pulsos média foi de 100%.

Com a posição de todos os pulsos no tempo, foi calculado a dinâmica em cada instante e a dinâmica média no intervalo de 4 segundos. A Figura 4-7 apresenta os resultados da dinâmica instantânea e média para os 7 vídeos. A linha azul corresponde a

variação da amplitude entre picos consecutivos e a linha verde é a média entre 100 amostras (4 segundos).

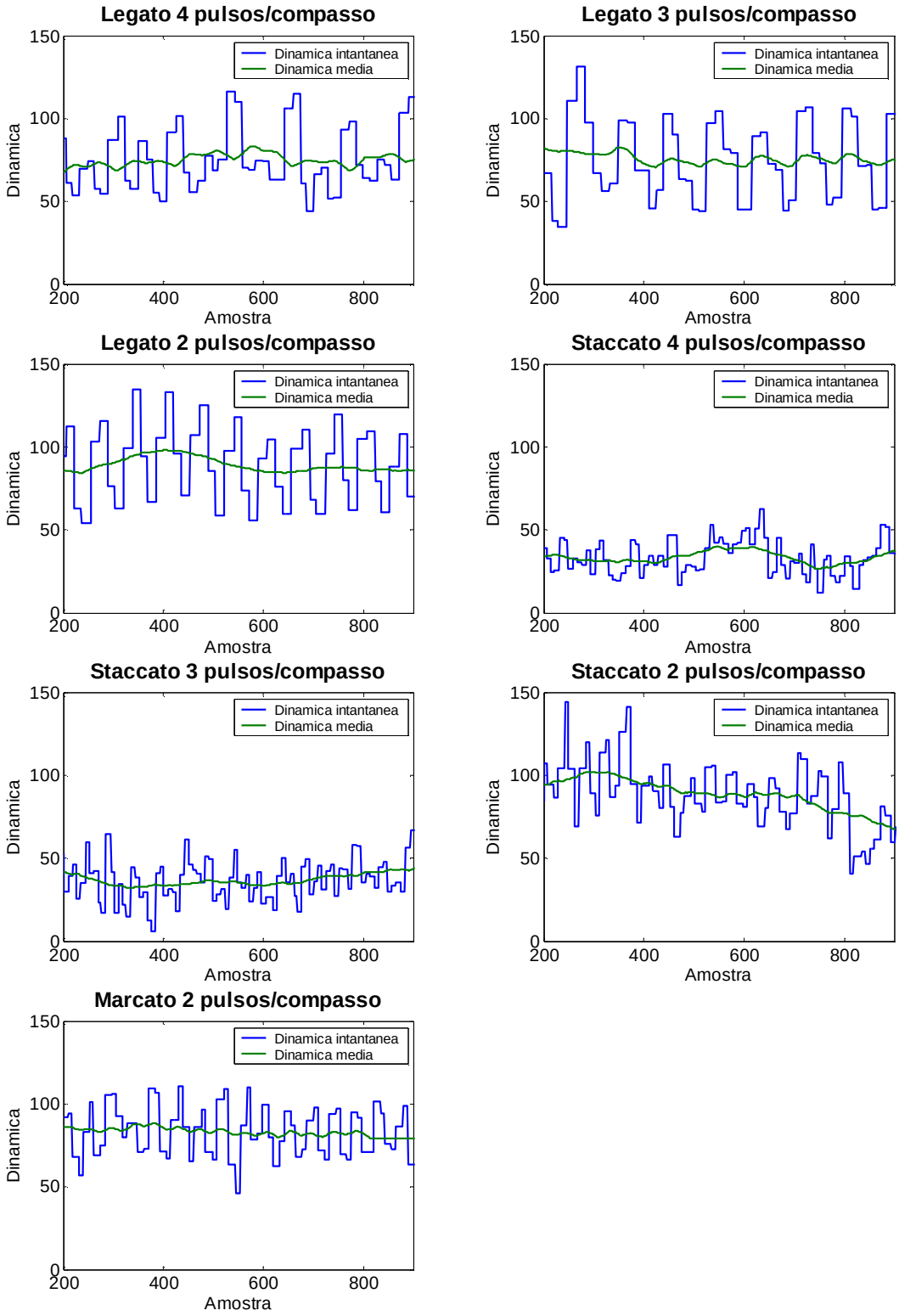


Figura 4-7: Dinâmica instantânea e média para os 7 vídeos.

A Figura 4-8 reúne os resultados da dinâmica média para os 7 vídeos.

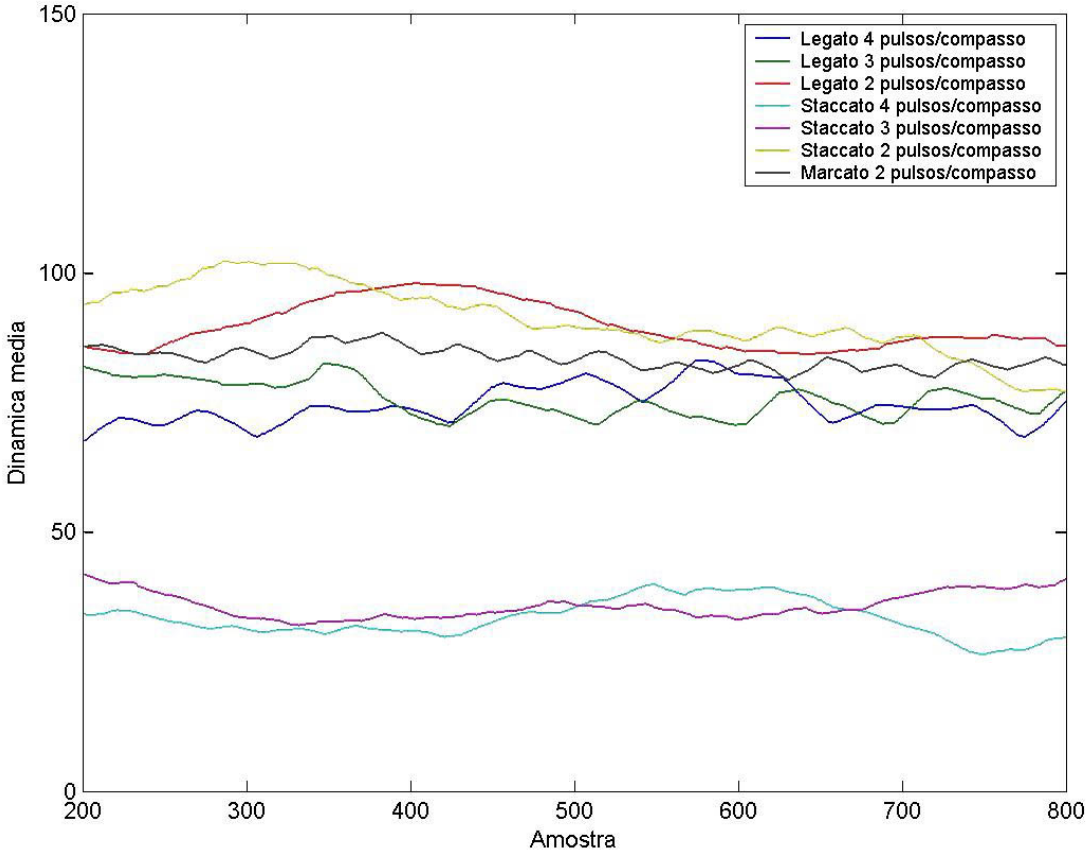


Figura 4-8: Dinâmica média nos 7 vídeos

Pode-se observar que as variações na amplitude não foram grandes, indicando que o maestro não desejava grandes variações na dinâmica nos 7 vídeos.

O objetivo do maestro na gravação dos vídeos, utilizados na pesquisa, era manter constância na amplitude da música. Os resultados obtidos nesta pesquisa condizem com o que o maestro desejava transmitir, e com o que um músico pode observar nos vídeos. A identificação da amplitude está diretamente relacionada a identificação do *tempo*, pois ambas dependem da identificação do pulso. Os bons resultados obtidos no reconhecimento do *tempo* e da amplitude indicam um bom reconhecimento dos pulsos.

4.2.4 *Gestos da mão direita*

Como visto anteriormente, os gestos da mão direita do maestro, geralmente, indicam a articulação, porém em alguns momentos podem indicar expressão. Para esta pesquisa foram separados 7 tipos de gestos de articulação e 3 gestos de expressão.

Para a análise da articulação foram utilizados os mesmos 7 vídeos usados para dinâmica e tempo. Cada vídeo contém 30 repetições do movimento. Os vídeos foram segmentados, sendo que 20 amostras foram utilizadas para treino das redes e 10 amostras para teste. A segmentação foi feita utilizando a identificação de pulsos citada no item 2.2.

Os vídeos representam os movimentos de *legato, staccato* e *marcato,* com variações na composição do compasso, com o número de pulsos entre 2 e 4. Os movimentos de expressão são: *gradual crescendo, gradual diminuendo* e sem dinâmica.

Os testes foram divididos em duas etapas. Na primeira os vídeos foram separados em 4 grupos:

- Expressão;
- Articulação com 4 pulsos/compasso;
- Articulação com 3 pulsos/compasso;
- Articulação com 2 pulsos/compasso.

Os grupos foram inicialmente treinados por redes independentes. Em uma segunda etapa foram feitos testes com todos os 10 tipos de gestos identificados por uma única rede.

Foram feitos testes utilizando as redes T-CombNET e Elman e os resultados obtidos foram comparados com os obtidos pela HMM [8]. Os resultados para cada grupo estão descritos nas Tabela 4-3 a Tabela 4-6.

Tabela 4-3: Taxa de reconhecimento para expressão da mão direita.

Movimento	Elman	T-CombNET	HMM
Legato* + *gradual crescendo	100%	100%	100%
Legato* + *gradual diminuendo	100%	70%	100%
***Legato* sem dinâmica**	100%	100%	100%
Média	**100%**	**90%**	**100%**

Tabela 4-4: Taxa de reconhecimento de gestos de articulação em movimentos de 4 pulsos/compasso.

	Movimento	Elman	T-CombNET	HMM
4 pulsos	*Legato*	100%	100%	100%
	Staccato	100%	100%	90%
	Média	**100%**	**100%**	**95%**

Tabela 4-5: Taxa de reconhecimento de gestos de articulação em movimentos de 3 pulsos/compasso.

	Movimento	Elman	T-CombNET	HMM
3 pulsos	*Legato*	100%	100%	100%
	Staccato	100%	100%	100%
	Média	**100%**	**100%**	**100%**

Tabela 4-6: Taxa de reconhecimento de gestos de articulação em movimentos de 2 pulsos/compasso

	Movimento	Elman	T-CombNET	HMM
2 pulsos	*Legato*	100%	100%	100%
	Staccato	100%	100%	100%
	Marcato	100%	100%	100%
	Média	**100%**	**100%**	**100%**

As estruturas das redes que obtiveram melhores resultados nos testes estão descritas na Tabela 4-7.

Um dos motivos que ajudou a obter as altas taxas de reconhecimento foi a simplificação feita ao separar manualmente os vídeos em 4 grupos, agrupados por tipo e número de pulso.

Na segunda etapa, os testes foram feitos com os 10 movimentos sendo identificados por uma única estrutura de rede, aproximando-se do reconhecimento real feito por um músico.

A rede Elman obteve melhor eficiência quando utilizou 180 neurônios na camada escondida. A rede T-CombNET obteve melhores resultados com a rede tronco sendo treinada pela segunda dimensão, e a rede tronco separou os vetores de entrada em 13 grupos. As configurações das redes estão descritas na última linha da Tabela 4-7, e os resultados obtidos na Tabela 4-8.

O tempo de treinamento da rede Elman foi de 40 horas utilizando um processador AMD Atlhon 64 3200 MHz com 512Mb de RAM, e o tempo para identificação de um vídeo foi em média de 4 segundos. A rede T-CombNET demorou 70 minutos para treinar e 8 segundos para validar.

Tabela 4-7: Estruturas das redes com melhores resultados nos testes.

Experiência	Estrutura da rede Elman	Estrutura da T-CombNET
Vídeos de 2 pulsos	Neurônios de entrada: 3 Neurônios camada escondida: 30 Neurônios de saída: 3	Normalização para LVQ: 150 Dimensões usadas na rede tronco: 1, 2 Dimensões usadas na rede galho: 1, 2, 3 Número de galhos 2 Galho 1: Neurônios de entrada: 3; Neurônios camada escondida: 30; Neurônios de saída: 3 Galho 2: Neurônios de entrada: 3; Neurônios camada escondida: 30; Neurônios de saída: 3
Vídeos de 3 pulsos	Neurônios de entrada: 3 Neurônios camada escondida: 20 Neurônios de saída: 2	Normalização para LVQ: 150 Dimensões usadas na rede tronco: 1, 2 Dimensões usadas na rede galho: 1, 2, 3 Número de galhos 2 Galho 1: Neurônios de entrada: 3; Neurônios camada escondida: 20; Neurônios de saída: 2 Galho 2: Neurônios de entrada: 3; Neurônios camada escondida: 20; Neurônios de saída: 2
Vídeos de 4 pulsos	Neurônios de entrada: 3 Neurônios camada escondida: 20 Neurônios de saída: 2	Normalização para LVQ: 150 Dimensões usadas na rede tronco: 1, 2 Dimensões usadas na rede galho: 1, 2, 3 Número de galhos 2 Galho 1: Neurônios de entrada: 3; Neurônios camada escondida: 20; Neurônios de saída: 2 Galho 2: Neurônios de entrada: 3; Neurônios camada escondida: 20; Neurônios de saída: 2
Vídeos de expressão	Neurônios de entrada: 3 Neurônios camada escondida: 30 Neurônios de saída: 3	Normalização para LVQ: 150 Dimensões usadas na rede tronco: 1, 2 Dimensões usadas na rede galho: 1, 2, 3 Número de galhos 2 Galho 1: Neurônios de entrada: 3; Neurônios camada escondida: 30; Neurônios de saída: 3 Galho 2: Neurônios de entrada: 3; Neurônios camada escondida: 30; Neurônios de saída: 3
Todos os gestos da mão direita juntos	Neurônios de entrada: 3 Neurônios camada escondida: 180 Neurônios de saída: 9	Normalização para LVQ: 100 Dimensões usadas na rede tronco: 1, 2 Dimensões usadas na rede galho: 1, 2, 3 Número de galhos 13 Galho 1,3,4,5,7,8,9,10,11,13: Neurônios de entrada: 3; Neurônios camada escondida: 20; Neurônios de saída: 1 Galho 2,6,12: Neurônios de entrada: 3; Neurônios camada escondida: 40; Neurônios de saída: 2

Tabela 4-8: Taxa de reconhecimento de gestos de articulação + expressão, com os 9 movimentos identificados pela mesma rede.

	Movimento	**Elman**		**T-CombNET**
		2D	**3D**	**3D**
4 pulsos	*Legato*	100%	100%	100%
	Staccato	100%	100%	90%
3 pulsos	*Legato*	100%	100%	90%
	Staccato	100%	100%	90%
2 pulsos	*Legato*	100%	100%	20%
	Staccato	100%	100%	80%
	Marcato	100%	100%	100%
Expressão	*Legato + gradual crescendo*	100%	100%	20%
	Legato + gradual diminuendo	100%	100%	80%
	Média	**100%**	**100%**	**81%**

A separação dos gestos em pequenos grupos fez com que ambas as redes neurais conseguissem resolver o problema com grande facilidade. A separação seria uma boa opção para facilitar o reconhecimento, porém não existe separação temporal e nem espacial na execução dos movimentos, eles são executados em seqüência pela mesma mão e podem ser executados a qualquer momento. Isso faz com que não tenha uma maneira fácil de fazer uma pré-seleção do gesto para um grupo. Por estes motivos a utilização de apenas uma rede neural para reconhecer todos os gestos aproxima-se mais do reconhecimento real, feito por um músico.

Nos teste feitos com apenas uma rede neural os resultados obtidos pela rede Elman foram de 100% de taxa de reconhecimento médio, contra 81% da rede T-CombNET. O tempo de treinamento foi considerado muito alto para a rede Elman, comparado ao

treinamento da rede T-CombNET. O tempo de identificação da T-CombNET foi o dobro da rede Elman.

Os resultados de 100% de taxa de reconhecimento incentivaram testes utilizando apenas duas dimensões. Utilizando a rede Elman alimentada apenas com as coordenadas da câmera frontal foi obtido o resultado de 100% de taxa de reconhecimento média. Este resultado demonstra que para o reconhecimento dos vídeos utilizados neste experimento não é necessária a utilização de uma segunda câmera.

A rede Elman mostrou-se mais eficaz no reconhecimento dos gestos do que a rede T-CombNET e do que HMM. Ressalta-se que para HMM não temos resultados do teste com todos os gestos juntos. Porém como o HMM já apresentou erros na identificação dos gestos separados em pequenos grupos, é esperado um pior desempenho no experimento mais complexo.

4.2.5 *Gestos da mão esquerda*

Os gestos da mão esquerda do maestro representam gestos de expressão. Para a análise de expressão foram utilizados 5 movimentos da mão esquerda. Para cada movimento foram gravados 30 vídeos, destes 20 foram utilizados para treinar as redes e 10 para testar.

Os vídeos da mão esquerda representam os movimentos *crescendo+corte, diminuendo+corte, fermata+click, accent* e *expansion*.

Assim como a análise de articulação, foram feito testes com as redes Elman e T-CombNET. As taxas de reconhecimento de ambas as redes para a mão esquerda são mostrados na Tabela 4-9 .

Tabela 4-9: Taxa de reconhecimento para expressão da mão esquerda.

Movimento	Elman	T-CombNET	HMM
Crescendo+corte	100%	70%	100%
Diminuendo+corte	100%	100%	100%
Fermata+click,	100%	90%	90%
Accent	100%	90%	100%
Expansion	100%	100%	100%
Média	**100%**	**88%**	**98%**

A rede Elman obteve melhores resultados com 100 neurônios na camada escondida. Para a rede T-CombNET foi obtida com a rede tronco sendo treinada com as dimensões 1 e 2, e as redes galhos com 20 ou 40 neurônios na camada escondida. Tais configurações estão descritas na Tabela 4-10.

Tabela 4-10: Configuração das redes.

<table>
<tr><th>Experiência</th><th>Estrutura da rede Elman</th><th>Estrutura da T-CombNET</th></tr>
<tr><td>Gestos da mão esquerda</td><td>Neurônios de entrada: 3
Neurônios camada escondida: 100
Neurônios de saída: 5</td><td>Normalização para LVQ: 100
Dimensões usadas na rede tronco: 1, 2
Dimensões usadas na rede galho: 1, 2, 3
Número de galhos 19
<table><tr><td>Galho 1,2,3,4,5,7,10,12,13,14,16,18,19
Neurônios de entrada: 3
Neurônios camada escondida: 20
Neurônios de saída: 1</td><td>Galho 6,8,9,11,15,17
Neurônios de entrada: 3
Neurônios camada escondida: 40
Neurônios de saída: 2</td></tr></table></td></tr>
</table>

Como os gestos de expressão são bem distintos os resultados com os 3 métodos foram significativos. A taxa de reconhecimento da rede Elman foi ligeiramente superior ao de HMM na identificação dos gestos de expressão. Por serem gestos curtos, o tempo de reconhecimento para todas as redes neurais foi rápido, podendo ambas serem utilizadas no reconhecimento em tempo real.

4.2.6 *Segmentação dos gestos*

A segmentação dos gestos é uma parte importante no reconhecimento de gestos. É necessário saber quando um gesto válido começa e termina, e quanto antes o gesto for identificado melhores serão os resultados.

A rede T-CombNET necessita conhecer o ponto de início e fim do movimento, para que seja feita a normalização dos dados para a rede tronco LVQ. Portanto, neste tipo de rede não é possível a identificação do gesto antes de que ele termine. A rede Elman não tem a necessidade de normalizar o sinal, portanto, pode identificar a presença de um gesto antes de que ele termine. Por este motivo a rede Elman é mais indicada para o reconhecimento em tempo real.

Os testes utilizando a rede Elman demonstram a grande capacidade deste tipo de rede de identificar a presença de um gesto. Na Figura 4-9 observa-se a saída da rede neural, quando excitada com um vídeo contendo os 9 gestos em seqüência. Cada cor indica a saída de um neurônio, a linha vertical pontilhada cinza indica a mudança de gesto. Pode se observar que em um pequeno período de tempo a saída referente ao gesto que está sendo executado passa a ser predominante, indicando a presença do gesto.

Um gesto é identificado como presente quando o somatório das 10 últimas saídas de um neurônio é a maior durante 10 quadros consecutivos.

O teste demonstrado na Figura 4-9 foi repetido 3000 vezes com diferentes combinações dos 9 vídeos. O resultado foi de que em média um gesto foi identificado quando 21% dele foi executado. Ou seja, um movimento *Legato* de 3 pulsos/compasso, que em média dura 80 quadros (3,2 segundos), é identificado em 16 quadros (0,64 segundos). No pior caso, um gesto foi identificado em 47% da execução do mesmo, ou seja, o sistema garante que o gesto será identificado antes da metade de sua execução.

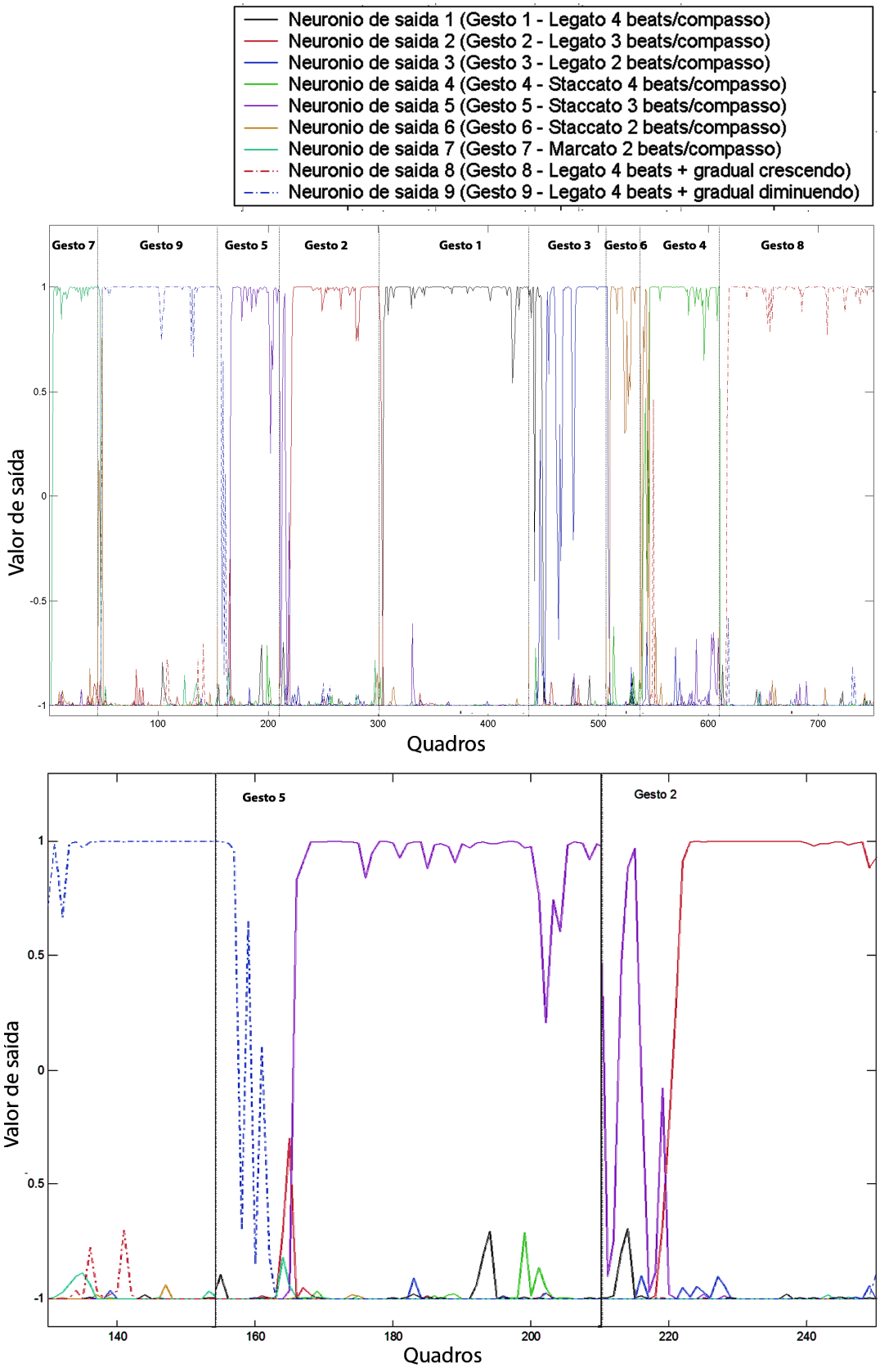


Figura 4-9: (a) Saída da rede neural para um vídeo com os 9 vídeos contínuos, (b) *Zoom* na transição do gesto 9 para 5 e do gesto 5 para o 2.

A identificação da transição entre gestos é um ponto importante no reconhecimento de gestos, e geralmente um forte empecilho para a implementação dos métodos em aplicações comerciais. Nesta pesquisa a rede Elman mostrou-se capaz de identificar a mudança de gestos sem a necessidade da intervenção do maestro para indicar começo e fim de um gesto, característica desejável e necessária para a utilização do método em tempo real.

4.2.7 *Testes complementares*

Para aferir os resultados do reconhecimento de gestos, os vídeos de treino e testes foram divididos em grupos, e este permutados para o treinamento e teste da rede. Como este teste é complexo e demanda tempo no treinamento, foi realizado apenas no experimento com os 9 gestos da mão direita, pois é o mais completo e complexo da pesquisa.

Os testes foram feitos apenas com a rede Elman. A configuração utilizada foi 3 neurônios na camada de entrada, 90 na escondida e 7 na de saída.

Os 30 vídeos de cada gesto (20 de treino e 10 de teste) foram separados em 3 grupos de 10 vídeos, de forma que todos os grupos contivessem o mesmo número de vídeos de um determinado gesto. Estes grupos foram permutados e 2 deles usados para treino e um para teste. O resultado está descrito na Tabela 4-11.

O resultado final foi inferior aos 100% do experimento retratado no item 4.2.4 . O tempo de treinamento das 3 redes para este teste foi de 72 horas, o que dificultou testes com diferentes configurações de rede, o que poderá elevar os resultados ao mesmo 100% obtido com diferente mistura dos vídeos.

Tabela 4-11: Taxa de reconhecimento de gestos de articulação + expressão (9 movimentos) com os vídeos separados em 3 grupos (A,B e C)

Grupo de Treino	**Grupo de teste**	**Taxa de reconhecimento**
A, B	C	95 %
A, C	B	92 %
B, C	A	94 %
Média		**94%**

Estatisticamente este experimento é mais consistente do que o teste feito apenas com um grupo de vídeos. O teste da rede com diversas combinações faz com que o resultado seja facilmente reproduzido pois não é específico para uma combinação de gestos de treino e teste.

Capítulo 5

Conclusões

Este trabalho apresentou uma nova metodologia para análise de gestos de maestro utilizando redes neurais parcialmente recorrentes. O sistema proposto diferentemente de outros trabalhos objetivou o reconhecimento de gestos de ambas as mãos, obtendo bons resultados.

Os testes realizados no reconhecimento dos gestos da mão esquerda e os da mão direita separando em pequenos grupos demonstraram uma facilidade na identificação, fazendo com que os testes não evidenciassem as diferença entre os métodos propostos e HMM. A união de todos os gestos da mão direita enriqueceu a pesquisa aproximando o sistema de reconhecimento do feito pelo músico.

A rede neural Elman mostrou-se mais precisa do que HMM no reconhecimento de gestos de maestro. HMM mostrou erros, ainda que poucos, em testes simples.

As diferenças entre as duas redes neurais testadas foram mais evidentes nos experimentos realizados nos gestos da mão direita, quando treinados por uma só rede. A rede Elman obteve melhores taxas de reconhecimento, porém seu tempo de treinamento foi cerca de 40 vezes maior. O tempo de reconhecimento da rede T-CombNET foi duas vezes maior que o da Elman, diferença que faz com que a T-CombNET não seja mais capaz de fazer a identificação em tempo real no computador usado na pesquisa.

Ainda que utilizássemos um computador com maior capacidade de processamento, a rede T-CombNET não se tornaria mais interessante para esta aplicação pois tem a desvantagem de necessitar de uma normalização no tempo para a

identificação dos gestos. Isso impossibilita que o gesto seja identificado antes da finalização do mesmo. Este problema não é crítico quando a aplicação não necessita de uma ação em tempo real, por exemplo, reconhecimento de linguagem de sinais, porém os gestos do maestro necessitam de uma reação quase instantânea na música.

O fato das taxas de reconhecimento obtidas pela rede T-CombNET serem menores do que as obtidas pela rede Elman pode ser atribuído a ineficiência do estágio de pré-seleção realizado pela rede tronco LVQ.

Os resultados obtidos fazem da rede Elman um futuro promissor no estudo dos gestos de maestro, pelo resultado obtido no reconhecimento de gestos e também pela facilidade na segmentação temporal.

Trabalhos futuros podem incrementar a pesquisa na expansão do banco de dados. Visando a independência do maestro, os mesmos gestos poderão ser gravados por outros maestros. O aumento do número de gestos, utilizando gestos personalizados do maestro também poderá ser abordado. Esse aumento no número de gestos, e conseqüentemente a dificuldade do reconhecimento dos mesmos, faz com que as diferença entre os métodos se amplie.

Estudos estatísticos, de custo computacional e a análise de diferentes métodos para o treinamento da rede também serão enriquecedores na pesquisa.

Este trabalho focou-se no reconhecimento dos gestos e não desenvolveu pesquisas na aplicação dos parâmetros retirados na música, assim como no reconhecimento em tempo real. Este aspecto também pode ser alvo de uma continuação da pesquisa.

Referências

[1] Buxton, W., W. Reeves, G. Fedorkov, K. C. Smith, and R. Baecker (1980). “A microprocessor-based conducting system”, Computer Music Journal 4(1), 8–21.

[2] Bertini, G. and P. Carosi (1992). “Light baton: A system for conducting computer music performance”, Proceedings of the International Computer Music Conference, pp. 73–76. International Computer Music Association.

[3] Borchers, J., W. Samminger, and M. Muhlhauser (2002). “Engineering a realistic real-time conducting system for the audio/video rendering of a real orchestra.”, Proceedings of the 4th International Symposium on Multimedia Software Engineering, pp. 352–362. International Computer Music Association.

[4] Garnett, G. E., M. Jonnalagadda, I. Elezovic, T. Johnson, and K. Small (2001). “Technological advances for conducting a virtual ensemble.”, Proceedings of the International Computer Music Conference, pp. 167–169. International Computer Music Association.

[5] Haflich, F. and M. Burns (1983). “Following a conductor: The engineering of an input device”, Proceedings of the International Computer Music Conference. International Computer Music Association.

[6] Keane, D. and P. Gross (1989). “The midi baton”, Proceedings of the International Computer Music Conference, pp. 151–154. International Computer Music Association.

[7] Lee, M., G. Garnett, and D. Wessel (1992). “An adaptive conductor follower”, Proceedings of the International Computer Music Conference, pp. 454–455. International Computer Music Association.

[8] Kolesnik, P. and Wanderley, M. “Recognition, analysis and performance with expressive conducting gestures”, Master Thesis, McGill University, Canadá, 2004.

[9] Kremer S.C., “On the Computational Power of Elman-Style Recurrent Networks”, IEEE Trans. on Neural Networks, Vol. 6, No. 4, 1995.

[10] Lamar, M. V. “Hand Getsure Recognition using TCombNET: A Neural Network dedicated to Temporal Information Processing”, Ph.D. Thesis, Nagoya Institute of Technology, Japan, 2001.

[11] Kohonen, T. “Improved Versions of Learning Vector Quantization”, International Joint Conference on Neural Networks, San Diego, 1990.

[12] Werbos, P. J. “ Backpropagation Through Time: What It Does and How to Do It”, Proceedings of IEEE, October 1990.

[13] Camurri, A., P. Coletta, M. Peri, M. Ricchetti, A. Ricci, R. Trocca, and G. Volpe. “A real-time platform for interactive performance.”, *Proceedings of the International Computer Music Conference*. International Computer Music Association, 2000.

[14] “EyesWeb - toward gesture and affect recognition in dance/music interactive systems”, Proc. IEEE Multimedia Sustems '99, Firenze, Italy, June 1999.

[15] McCulloch, W. S. and Pitts, W. “A Logical Calculus of the Ideas Immanent in Nervous Activity”, Bulletin of Mathematical Biophysics, Vol5, pp.115-133, 1943.

[16] Hopfield, J. J. and Tank, D. W. “Computing with Neural Circuits: A Model”, Science, Vol. 233, Aug. 1986.

[17] Williams, R. J. and Zipser, D. “A Learning Algorithm for Continually Running Fully Recurrent Neural Networks”, ICS Report 8805, Oct. 1988.

[18] Rudolph, M. “The Grammar of Conducting: A comprehensive guide to baton technique and interpretation.”, Toronto: Maxwell Macmillan Canada (1994).

[19] Mathews, M. V. “The Conductor program”, Proceedings of the International Computer Music Conference, Cambridge, Massachusetts (1976).

[20] Mathews, M.. “Current Directions in Computer Music Research”. MIT Press (1989).

[21] Morita, H., S. Otheru, and S. Hashimoto. “Computer music system that follows a human conductor.” Proceedings of the International Computer Music Conference, pp. 207–210. International Computer Music Association (1989).

[22] Morita, H., S. Otheru, and S. Hashimoto. “Knowledge information processing in conducting computer music performance.” Proceedings of the International Computer Music Conference, pp. 332–334. International Computer Music Association (1990).

[23] Mathews, M. V. “The Radio Baton and the Conductor Program, or:Pitch—the most important and least expressive part of music.” Computer Music Journal Vol.15 , 37–46 (1991).

[24] Tobey, F. and I. Fujinaga. “Extraction of conducting gestures in 3d space.” Proceedings of the International Computer Music Conference, pp. 305–307. International Computer Music Association (1996).

[25] Usa, S. and Y. Mochida. “A conducting recognition system on the model of musicians process.” Journal of Acoustical Society of Japan Vol.19, 275–287 (1998).

[26] Marrin, T. and R. Picard. “The Conductors Jacket: A device for recording expressive musical gestures.” Proceedings of the International Computer Music Conference, pp. 215–219. International Computer Music Association (1998).

[27] Ilmonen, T. and T. Takala. “Conductor following with Artificial Neural Networks.” Proceedings of the International Computer Music Conference, pp. 367–370. International Computer Music Association (1999).

[28] Segen, J., A. Mujumder, and J. Gluckman. “Virtual dance and music conducted by a human conductor”. Eurographics Vol.19 (2000).

[29] Murphy, D., T. H. Andersen, and K. Jensen. “Conducting audio files via Computer Vision.” Proceedings of the 2003 International Gesture Workshop, Genoa, Italy (2003)

[30] Tamura, S. and Kawasaki, S. “Recognition of Sign Language Motion Images”, Pattern Recognition,

[31] Takahashi, T. and Kishino, F. “Hand Gesture Coding Based on Experiments using a Hand Gesture Interface Device”, SIGCHI Bulletin, 23(2), pp. 67-73, April, 1991.

[32] Murakami, K. and Taguchi, H. “Gesture Recognition using Recurrent Neural Networks” *CHI’91 Conference Proceedings*, pp. 237-241, 1991.

[33] Iwata, A., Suwa, Y., Ino, Y., Hotta, K. I. and Suzumura, N. “Hand-Written Japanese Kanji character recognition by a structured self-growing neural network”, *Artificial Neural Networks*, Vol.2, I. Aleksander adn J. Taylor (editors), Elsevier Science Publishers, pp.1189-1192, 1992.

[34] Waibel, A. T., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K. J. “Phoneme Recognition using Time-Delay Neural Networks”,*IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol.37, 1989.

[35] Fransconi, P., Gori, M. and Soda, G. “Local Feedback Multilayered Networks”, *Neural Computation*, Vol.4, pp.120-130, 1992.

[36] Choi, H. I. and Rhee, P. K. “Hand gesture recognition using HMMs”, *Expert System with applications*, Volume 17, October 1999.

[37] Forney, G.D. “The Viterbi Algorithm”, *Procs of the IEEE*, 1973.

[38] Lou H.L. “Implementing the Viterbi Algorithm”, *IEEE Signal Processing Magazine*, 1995.

[39] Baum, L. E., Peterie, T., Souled, G. and Weiss, N. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, no. 1, 1970.

[40] Elman, J.L. “Finding Structure in Time.” *Cognitive Science* vol.14, 199