UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Dhiego Cristiano Oliveira da Silva Sad

Um descritor tensorial de movimento baseado em múltiplos estimadores de gradiente

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Marcelo Bernardes Vieira

Juiz de Fora

2013

Dhiego Cristiano Oliveira da Silva Sad

Um descritor tensorial de movimento baseado em múltiplos estimadores de gradiente

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Aprovada em 22 de Fevereiro de 2013.

BANCA EXAMINADORA

Prof. D.Sc. Marcelo Bernardes Vieira - Orientador
Universidade Federal de Juiz de Fora

Prof. D.Sc. Rodrigo Luis de Souza da Silva
Universidade Federal de Juiz de Fora

Prof. D.Sc. Antônio Alberto Fernandes de Oliveira
Universidade Federal do Rio de Janeiro

Aos meus pais, namorada e
amigos pelo apoio incondicional.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais e à Karoline, minha namorada e eterno amor, pelo total apoio e dedicação em todos os passos desta caminhada. Aos meus colegas do Grupo de Computação Gráfica, Imagem e Visão por colaborarem no desenvolvimento do método proposto neste trabalho. Finalmente, agradeço à CAPES pelo auxílio financeiro.

"A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê."

(Arthur Schopenhauer)

RESUMO

Este trabalho apresenta uma nova abordagem para a descrição de movimento em vídeos usando múltiplos filtros passa-banda que agem como estimadores derivativos de primeira ordem. A resposta dos filtros em cada quadro do vídeo é extraída e codificada em histogramas de gradientes para reduzir a sua dimensionalidade. Essa combinação é realizada através de tensores de orientação. O grande diferencial deste trabalho em relação à maioria das abordagens encontradas na literatura é que nenhuma característica local é extraída e nenhum método de aprendizagem é realizado previamente, isto é, o descritor depende unicamente do vídeo de entrada. Para o problema de reconhecimento da ação humana utilizando a base de dados KTH, nosso descritor alcançou a taxa de reconhecimento de 93,3% usando três filtros da família Daubechies combinado com mais um filtro extra que é a correlação entre esses três filtros. O descritor resultante é então classificado através do SVM utilizando um protocolo *two-fold*. Essa classificação se mostra superior para a maioria das abordagens que usam descritores globais e pode ser comparável aos métodos do estado-da-arte.

Palavras-chave: Múltiplos filtros. Descritor de movimento. Filtros correlacionados. Tensor de orientação. Reconhecimento de ações humanas.

ABSTRACT

This work presents a novel approach for motion description in videos using multiple band-pass filters that act as first order derivative estimators. The filters response on each frame are coded into individual histograms of gradients to reduce their dimensionality. They are combined using orientation tensors. No local features are extracted and no learning is performed, i.e., the descriptor depends uniquely on the input video. Motion description can be enhanced even using multiple filters with similar or overlapping frequency response. For the problem of human action recognition using the KTH database, our descriptor achieved the recognition rate of 93,3% using three Daubechies filters, one extra filter designed to correlate them, two-fold protocol and a SVM classifier. It is superior to most global descriptor approaches and fairly comparable to the state-of-the-art methods.

Keywords: Multifilter analysis. Motion descriptor. Correlation filter. Orientation tensor. Human action recognition.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

No final da década de 1970 surgiram as primeiras pesquisas voltadas para a área da visão computacional, sendo definida como um conjunto de métodos e técnicas através dos quais sistemas artificiais são capazes de obterem informações de imagens ou quaisquer dados multi-dimensionais. Um sistema de visão completo pode ser dividido da seguinte forma (MARR et al., 2010):

- Aquisição de Imagem: consiste em obter uma sequência de imagens digitais através de sensores geralmente contidos em câmeras digitais, como por exemplo, webcam. Dependendo to tipo de sensor o resultado da captação pode variar entre uma imagem bidimensional ou em uma sequência de imagens. Os pixels indicam em cada coordenada valores de intensidade de luz em uma cor.

- Pré-processamento: consiste em aplicar métodos de processamento de imagem, por exemplo, filtros de suavização, para reduzir os ruídos gerados pela aquisição da imagem antes de extrair informações.

- Extração de características: consiste em capturar informações de uma imagem. Uma imagem é formada por modelos matemáticos, como por exemplo matrizes, estas contêm características que podem matematicamente ser identificadas como: textura, bordas e etc.

- Detecção e segmentação: consiste em destacar uma determinada região de uma imagem e segmentá-la, com a finalidade de guardar essa informação para processamento posterior.

- Pós-processamento: consiste na verificação dos dados, a estimativa de parâmetros sobre a imagem e a classificação dos objetos detectados em diferentes categorias.

O foco de estudo deste trabalho, que se insere na área de visão computacional, está no reconhecimento de movimentos em vídeos. Movimento é a principal característica que representa a informação semântica em vídeos. Detectar um objeto ou uma pessoa e rastreá-lo é de grande interesse em diversas aplicações de segurança, como por exemplo rastreamento de mísseis e detecção de movimento em sistemas de vigilância.

Este trabalho utiliza uma combinação de filtros para extrair diferentes espectros do vídeo. As respostas dos filtros em cada quadro do vídeo são extraídas e codificadas em histogramas de gradientes (ZELNIK-MANOR; IRANI, 2001) para redução de dimensionalidade, ou seja, conseguir de forma condensada representar toda informação de movimento extraída dos vídeos. Esses filtros agem como operadores derivativos para extração de atributos locais de cada pixel. O gradiente obtido representa a máxima variação da intensidade de briho em um ponto da imagem. Com isso, é possível armazenar essas informações em descritores. Os vídeos utilizados neste trabalho são oriundos da base de dados KTH (SCHULDT et al., 2004).

1.1 DEFINIÇÃO DO PROBLEMA

O principal problema deste trabalho é encontrar a melhor correlação de filtros derivativos para extração de informações de movimento em vídeos. Dessa forma pode-se analisar diferentes porções do espectro de cada vídeo, aumentando assim a quantidade de informação de movimento capturada em cada filtragem.

1.2 OBJETIVOS

O objetivo primário deste trabalho é investigar e propor uma combinação de filtros que agem como estimadores derivativos para representar movimentos em vídeos.

Como objetivo secundário, deve-se obter um descritor que represente de forma compacta toda informação capturada para um dado vídeo.

1.3 CONTRIBUIÇÕES E PUBLICAÇÕES

Este trabalho é uma continuação de duas dissertações (MOTA, 2011; PEREZ, 2012) de mestrado e um artigo (PEREZ et al., 2012), cujo objetivo é estender os trabalhos anteriores, visando um resultado melhor no que diz respeito à precisão no reconhecimento de ações em vídeos.

Em Mota (2011) propõe-se um descritor global de movimento baseado em um tensor de orientação. Este descritor, assim como em Kihl et al. (2010), também é extraído da projeção do fluxo óptico em uma base ortogonal de polinômios. Neste trabalho, tensores são usados como acumuladores de informação de movimento.

No trabalho de Perez et al. (2012) é realizada uma combinação entre tensores de segunda ordem e histogramas de gradientes na geração dos descritores utilizando informação de todo quadro, sendo mais simples e menos custoso computacionalmente. Histogramas de gradiente foram usados como redutores de dimensionalidade do gradiente calculado.

A principal contribuição deste trabalho é um novo método para construção de um descritor global de movimento baseado na aplicação de múltiplos filtros no vídeo. Usando um classificador SVM, nosso descritor alcança taxas de reconhecimento (93,3%) que podem ser comparadas ao estado-da-arte e superior aos descritores globais encontrados na literatura.

Este trabalho gerou uma submissão no *International Conference on Image Processing* (ICIP) 2013 intitulada *A tensor motion descriptor based on multiple gradient estimators*

1.4 TRABALHOS RELACIONADOS

Neste capítulo são apresentados trabalhos relacionados à criação de descritores de movimento. Alguns métodos presentes na literatura utilizam técnicas distintas tanto para a análise do vídeo no domínio espacial, quanto no domínio da frequência.

1.4.1 BASE DE DADOS

O conjunto de dados KTH (SCHULDT et al., 2004) é considerado a base de dados mais amplamente utilizada para o reconhecimento da ação humana. Essa base de dados foi introduzida por Schuldt et al. e contém seis tipos de ações humanas (caminhar, correr, trotar, boxe, acenando com a mão e mão batendo palmas), que são executadas por 25 atores em quatro cenários diferentes. Todas as 2391 sequências têm uma resolução espacial de 160x120 pixels, uma taxa de frames de 25 quadros por segundo e cerca de 4 segundos de duração. O fundo é estático com alguns movimentos de câmera (Fig 1.1).

1.4.2 DESCRITORES LOCAIS

Para o problema de reconhecimento de ações humanas, diversos autores utilizam métodos para a criação de descritores locais. Entre eles, destacam-se aqueles que utilizam informações locais para extrair um maior número de características (LAPTEV et al., 2008). Em geral, os autores tentam combinar essas informações locais para obter uma melhor taxa de reconhecimento.

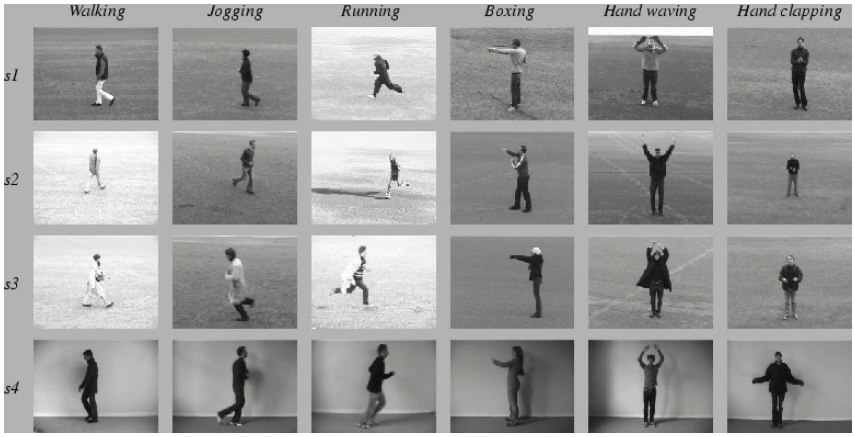


Figura 1.1: Base de dados KTH (SCHULDT et al., 2004).

Laptev et al. (2008) propõe um novo método para classificar movimentos em vídeos que é uma extensão de algumas técnicas conhecidas de reconhecimento em imagens para o domínio espaço-temporal. Para caracterizar o movimento, ele calcula histogramas em volumes espaço-temporais na vizinhança de pontos de interesse. Cada volume é subdividido em um conjunto de cubóides e para cada cubóide calculam-se histogramas de gradientes (HOG) e de fluxo óptico (HOF - *Histogram of Optical Flow*). Finalmente, esses descritores são normalizados e concatenados em um descritor. O conjunto desses descritores é chamado de *bag-of-visual-features* (BoF) e são utilizados para fazer uma posterior classificação dos vídeos.

Histogramas de gradientes orientados, são histogramas gerados a partir dos gradientes de imagens. Proposto inicialmente em Dalal e Triggs (2005) para a detecção humana em imagens, foi posteriormente estendido para o reconhecimento de ações em vídeos. Em Kläser et al. (2008) é proposto um descritor HOG em três dimensões (HOG3D) utilizando também a informação temporal do vídeo, além da informação espacial de cada quadro.

1.4.2.1 Descritores locais baseados em tensores

Tensores são poderosas ferramentas matemáticas que têm sido muito utilizadas nos últimos anos em diversas aplicações. No campo de reconhecimento de movimentos, poucos trabalhos utilizam tensor como um descritor para o reconhecimento de ações humanas. Os trabalhos que fazem uso de tensores podem ser classificados em dois tipos: os que utilizam operações tensoriais para ajudar na análise do vídeo (KIM et al., 2007; KRAUSZ; BAUCKHAGE, 2010) e aqueles que usam as propriedades do tensor, usando-o assim como

um descritor (KIHL et al., 2010; KHADEM; RAJAN, 2009).

1.4.2.2 Descritores locais baseados em banco de filtros

Técnicas que transformam o domínio são amplamente utilizadas no campo de processamento de imagem, tais como compressão e segmentação de imagens.

Em Shao e Gao (2010) é proposto um método para criação de descritores baseados em transformada wavelet. Inicialmente, os pontos de interesse são detectados. Em seguida, são extraídos cubóides em torno desses pontos. Para criar o descritor, são aplicadas wavelets Daubechies dentro desses cubóides a fim de obter as informações contidas em cada um deles. Finalmente, na fase de classificação, é utilizado um *SVM* com função kernel de base radial (*RBF*).

Em Minhas et al. (2010) é apresentado uma combinação de características espaço-temporais e características locais estáticas. Para determinar as características espaço-temporais, os coeficientes da wavelet complexa em diferentes sub-bandas são representadas por vetores de baixa dimensão. A transformada da wavelet complexa dual-tree (DT-CWT) é construída através de um par, ortogonal ou bi-ortogonal de bancos de filtros que trabalham em paralelo. Para determinar as características locais estáticas, foi utilizado o método conhecido como *Scale Invariante Feature Transform* (SIFT).

1.4.3 DESCRITORES GLOBAIS

Neste trabalho é possível observar que a utilização de descritores locais para o reconhecimento de ações humanas são mais explorados por alcançarem maiores taxas de reconhecimento. Porém, existe uma outra linha de pesquisa voltada para criação de descritores globais. Esses descritores, apesar de ainda não apresentarem uma taxa de reconhecimento superior a todos os descritores locais, conseguem atingir um determinado nível de simplicidade e robustez que proporciona uma classificação para o reconhecimento de ações humanas de forma rápida e independente das bases de vídeos utilizadas.

Um descritor global baseado em histograma de gradientes orientados (*HOG*) é apresentado em Zelnik-manor e Irani (2001). Esse descritor é aplicado utilizando a base de dados Weizmann (GORELICK et al., 2005). Para obter o descritor, são extraídas várias escalas temporais, através da construção de uma pirâmide temporal. Para calcular esta pirâmide, é aplicado um filtro passa-baixa em cada quadro do vídeo. Para cada escala, a

intensidade de cada pixel do gradiente é calculada. Em seguida, é criado um HOG para cada vídeo. Por fim, é realizado uma comparação com outros histogramas para classificar o banco de dados.

Utilizando a base de dados KTH, Laptev et al. (2007) estendeu o trabalho proposto em Zelnik-manor e Irani (2001) para criar um descritor global que pode ser aplicado de duas maneiras: a primeira é utilizando múltiplas escalas temporais como o original e o segundo é utilizando múltiplas escalas temporais e espaciais.

Solmaz et al. (2012) apresenta um descritor global baseado em um banco de 68 filtros de Gabor. Para cada vídeo, são extraídos vários quadros do vídeo e então é computado a Transformada Discreta de Fourier 3-D. Em seguida é feita a aplicação de cada filtro separadamente para o espectro de frequências, quantificando a produção de sub-volumes fixos. Em seguida, os resultados são concatenados e é realizada uma redução de dimensão através de uma técnica chamada Análise de Componentes Principais. Por fim é realizada uma classificação por SVM.

1.4.3.1 Descritores globais baseados em tensores

Em Mota (2011) é proposto um descritor global de movimento baseado em tensores de orientação. Esse tensor, assim como em Kihl et al. (2010), também são extraídos da projeção do fluxo óptico em uma base ortogonal de polinômios.

No trabalho de Perez et al. (2012) é realizada uma combinação entre tensores de segunda ordem e histogramas de gradientes na geração dos descritores utilizando informação de todo quadro, sendo mais simples e menos custoso computacionalmente.

Nesta dissertação, ao invés de usar apenas um filtro derivativo para extrair movimento, é utilizado uma combinação entre múltiplos filtros com intuito de extrair distintas características de movimento em cada vídeo.

2 FUNDAMENTOS

Neste capítulo são apresentados os conceitos básicos necessários para compreensão de cada etapa, essenciais para construção de um descritor para reconhecimento de ações humanas em vídeos.

2.1 SINAIS

Um sinal é uma função que representa uma quantidade física ou uma variável, contendo a informação acerca do comportamento ou natureza do fenômeno. Matematicamente podemos definir um sinal unidimensional como uma função de tempo $x(t)$. Se a variável t que representa o tempo mudar continuamente, então temos um sinal analógico ou contínuo (Fig. 2.1). Porém, se t for uma variável discreta, onde $x(t)$ só está definido em alguns pontos, temos então um sinal digital ou discreto (Fig. 2.2).

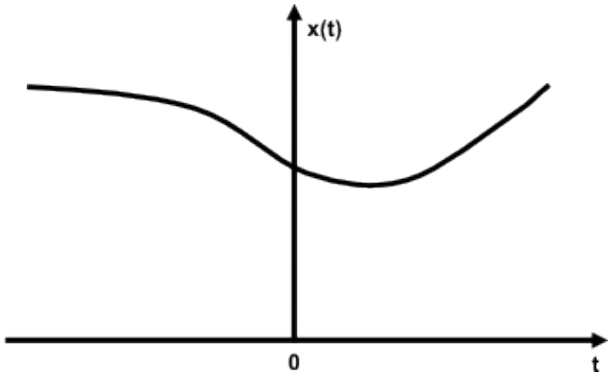


Figura 2.1: Representação de um sinal analógico.

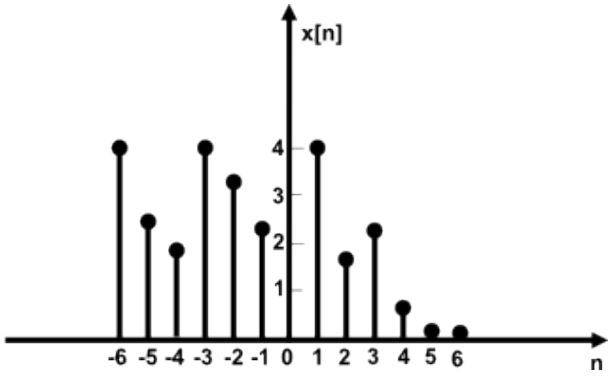


Figura 2.2: Representação de um sinal digital.

2.1.1 SINAIS DISCRETOS

Um sinal discreto é uma sequência de números indicados como $x[n]$, em que n é dito ser o índice de tempo, e $x[n]$ indica o valor do n-ésimo termo da sequência.

Cada termo da sequência $x[n]$ é também chamado de valor da amostra e pode assumir qualquer valor em um intervalo $x_{min} \leq x[n] \leq x_{max}$, e a variável n é chamada de índice da amostra.

Sinais discretos podem ser definidos somente para valores inteiros de n dentro de um intervalo $N_1 \leq n \leq N_2$. Podemos definir o tamanho da sequência $x[n]$ como $N \leq N_2 - N_1 + 1$. A sequência $x[n]$ é uma sequência finita se N é finito, caso contrário, $x[n]$ é uma sequência de tamanho infinito. Para efeitos de análise, é útil para representar os sinais como a combinação de sequências básicas (MILIC, 2009).

2.1.2 SISTEMAS DE SINAIS DISCRETOS

Um sistema discreto, é um algoritmo ou dispositivo físico que converte uma sequência de entrada para uma outra sequência de saída (MILIC, 2009). A relação de entrada-saída do sistema pode ser expressa matematicamente como:

$$y[n] = \Phi(x[n]), \tag{2.1}$$

onde o operador Φ representa a regra de uso para produzir o sinal de saída $y[n]$ a partir do sinal de entrada $x[n]$. Um sistema discreto é estável se qualquer sequência de entrada limitada produz uma sequência de saída limitada. Apenas os sistemas estáveis são de interesse prático. Um sistema discreto é causal se a saída depende apenas dos valores atuais e anteriores do sinal de entrada. Se $y[n_0]$ é a saída para o tempo de índice n, então $y[n_0]$ depende somente da amostra de entrada $x[n]$ para valores $n \leq n_0$.

2.2 SISTEMAS LINEARES E INVARIANTES NO TEMPO

Linear time-invariant (LTI) são sistemas lineares estáveis com o tempo invariante. A resposta do sistema para uma sequência de amostras unitárias $\delta[n]$ é chamada de resposta de impulso e é indicado por $h[n]$,

$$h[n] = \Phi(\delta[n]), \tag{2.2}$$

onde

$$\delta[n] = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} . \tag{2.3}$$

Um sistema LTI só é caracterizado por $h[n]$ se a sequência da saída do sistema pode ser representada como uma convolução da sequência de entrada e a resposta do impulso do sistema:

$$y[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot h[n-k]. \tag{2.4}$$

Essa convolução pode ser representada compactamente por

$$y[n] = x[n] * h[n]. \tag{2.5}$$

Um sistema LTI é considerado estável se o impulso de resposta satisfaz a seguinte condição:

$$\sum_{n=-\infty}^{\infty} \mid h[n] \mid < \infty. \tag{2.6}$$

Um sistema LTI é considerado causal se o impulso de resposta $h[n]$ é uma sequência causal dada por:

$$h[n] = 0, \text{ para } n < 0. \tag{2.7}$$

Um sistema LTI é considerado anti-causal se o impulso de resposta $h[n]$ é uma sequência anti-causal,

$$h[n] = 0, \text{ para } n > 0. \tag{2.8}$$

Um sistema LTI pode ser divido em duas categorias, uma é o sistema de resposta de impulso finito (FIR - *Finite Impulse Response*) a outra é o sistema de resposta de impulso infinito (IIR - *Infinite Impulse Response*).

Para um sistema FIR, $h[n]$ é de comprimento finito e a relação de entrada-saída é expressa como uma convolução de soma finita.

Para um sistema IIR, $h[n]$ é de comprimento infinito e a relação de entrada-saída é expressa como uma convolução de soma infinita.

2.2.1 FILTROS

Filtros são operadores essenciais para analisar, codificar e reconstruir sinais. Filtrar é um processo no qual as amplitudes da frequência de um sinal são alteradas ou até mesmo eliminadas. Neste trabalho a palavra *filtro* é utilizada para representar sistemas que fazem seleção de frequências. Sistemas LTI funcionam como um filtro a medida que o espectro do sinal de saída é igual ao sinal de entrada multiplicado pela resposta de impulso do sistema.

Um filtro ideal para seleção de frequência, é um filtro capaz de deixar passar determinado conjunto de frequências (banda de passagem) e rejeitar as demais (banda de corte).

1. **Filtro Passa Baixa ideal:**

   Um filtro passa baixa ideal pode ser representado pela seguinte expressão:

$$| H(f) | = \begin{cases} 1, & | f | < f_c \\ 0, & | f | > f_c \end{cases},$$

   conforme mostra a Figura 2.3.

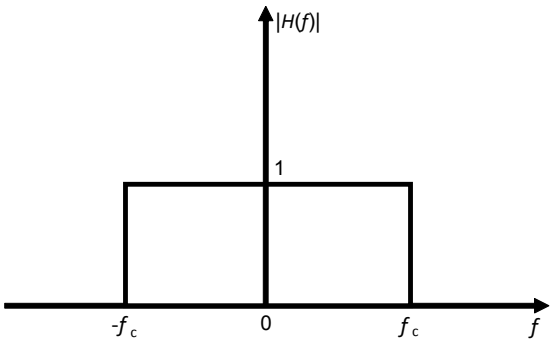


Figura 2.3: Magnitude da resposta de um filtro passa baixa ideal.

2. **Filtro Passa Alta ideal:**

   Um filtro passa alta ideal pode ser representado pela seguinte expressão:

$$| H(f) | = \begin{cases} 0, & | f | < f_c \\ 1, & | f | > f_c \end{cases},$$

   conforme mostra a Figura 2.4.

3. **Filtro Passa Banda ideal:**

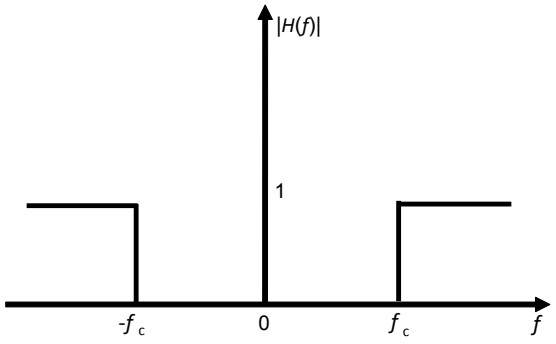


Figura 2.4: Magnitude da resposta de um filtro passa alta ideal.

Um filtro passa banda ideal pode ser representado pela seguinte expressão:

$$| H(f) |= \begin{cases} 1, & f_1 <| f |< f_2 \\ 0, & \text{caso contrário} \end{cases},$$

conforme mostra a Figura 2.5.

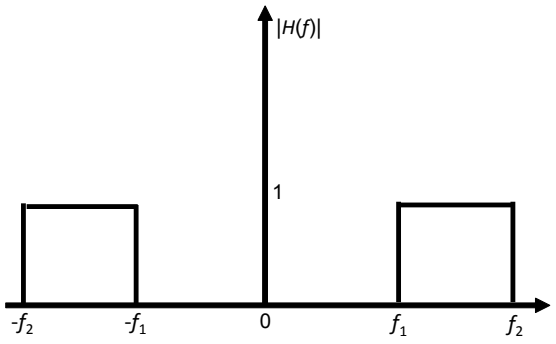


Figura 2.5: Magnitude da resposta de um filtro passa banda ideal.

2.2.2 FILTROS MULTITAXA

Sistemas lineares e invariantes no tempo (LTI) operam a uma taxa de amostragem única, a mesma na entrada e na saída do sistema, e em todos os nós no interior do sistema. Sistemas que utilizam taxas de amostragem distintas em diferentes etapas são chamados de sistemas multitaxa ou, neste caso, filtros multitaxa.

Os filtros multitaxa são usados para converter a taxa de amostragem dos dados de entrada para uma taxa de amostragem pretendida nos dados de saída, fornecendo diferentes taxas de amostragem sem destruir as componentes de sinal de interesse.

Os principais operadores multitaxa são os decimadores e os expansores, que operam em conjunto com filtros digitais, formando as estruturas de filtragem digital multitaxa.

Estas estruturas se combinam e formam os bancos de filtros digitais.

2.2.2.1 Operadores de decimação

A Figura 2.6 nos mostra o operador de decimação, também conhecido como *Down-Sampler* ou redutor de amostragem.

Figura 2.6: Bloco Operador de decimação por D.

Dada uma sequência de entrada pelo vetor $x[n]$, a sequência de saída é representada pelo vetor $y[m]$, de acordo com a Equação 2.9:

$$y[m] = x[D.n], \tag{2.9}$$

onde, D é um número inteiro. Apenas as amostras de $x[n]$ em que n é múltiplo de D são utilizadas pelo decimador. Por exemplo, se um conjunto de amostras for decimado por 2, a saída será gerada apenas com os valores de $x[n]$ para n par, ou n ímpar. Assim, terá a metade do número de amostras da sequência original, ou seja, a taxa de amostragem fica reduzida a metade.

Após a decimação, o espectro do sinal no domínio da frequência se alarga, podendo ocorrer superposição ou "*aliasing*". Este fenômeno ocorre quando o espectro do sinal original é maior que π/D. Quando ocorre a superposição, informações do sinal são perdidas, o que pode impossibilitar a sua reconstrução. Assim, o operador decimador é usualmente antecedido por filtro chamado de "*anti-aliasing*", para garantir que não ocorra superposição. Em geral, estes filtros são passa-baixa, com ganho unitário e frequência de corte em π/D.

2.2.2.2 Operadores de expansão

A Figura 2.7. nos mostra o operador de expansão, também conhecido como *Up-Sampler* ou expansor de amostragem.

Aplicando-se o operador de expansão no sinal discreto $x[n]$, será produzido amostras do sinal $y[m]$:

Figura 2.7: Bloco Operador de expansão por E.

$$y[m] = \begin{cases} x[n/E], & n = 0, \pm E, \pm 2E, \dots \\ 0, & n \neq 0, \pm E, \pm 2E, \dots \end{cases},$$

onde E é um número inteiro. O expansor produzirá na saída uma réplica de $x[n]$, se n é múltiplo de E, caso contrário a saída gerada possui valor zero. Um filtro passa baixa normalmente é utilizado depois do expansor, evitando que o espectro de frequência tenha imagens replicadas do espectro original. Um filtro passa-baixa com ganho E e frequência de corte em π/E normalmente é utilizado após o expansor para eliminar estas imagens, de maneira que o sinal volte a ter o mesmo espectro original, apenas com taxa de amostragem E vezes maior.

2.3 TENSOR DE ORIENTAÇÃO

Tensores estendem o conceito de vetores e matrizes para ordens maiores. Na terminologia tensorial, vetores são tensores de primeira ordem e matrizes são tensores de segunda ordem (WESTIN, 1994). Um tensor de orientação pode ser definido matematicamente como uma matriz real e simétrica para sinais m-dimensionais. Assim existem matrizes $n \times n$,

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix} \quad \text{e} \quad \mathbf{P} = [U_1 \; U_2 \; \dots \; U_n]$$

com $P^{-1} = P^t$ (ortogonal), tais que

$$\mathbf{T} = PDP^t. \tag{2.10}$$

ou seja,

$$\mathbf{T} = [U_1\ U_2\ \ldots\ U_n] \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_n \end{bmatrix} \begin{bmatrix} U_1^t \\ U_2^t \\ \vdots \\ U_n^t \end{bmatrix}$$

$$= [\lambda_1 U_1\ \lambda_2 U_2\ \ldots\ \lambda_n U_n] \begin{bmatrix} U_1^t \\ U_2^t \\ \vdots \\ U_n^t \end{bmatrix}$$

$$\mathbf{T} = \sum_{i=1}^{n} \lambda_i U_i U_i^t, \tag{2.11}$$

onde λ_i são os autovalores e U_i os respectivos autovetores.

Com o tensor de orientação, podemos representar as orientações em um campo de gradientes. Estes tensores são normalmente utilizados em aplicações da área de processamento de imagens e visão computacional para detecção de pontos de interesse.

2.4 HISTOGRAMA DE GRADIENTES

Neste trabalho os descritores são calculados de modo semelhante a Perez et al. (2012). O gradiente do j-ésimo quadro de um vídeo em um ponto p é dado por:

$$\vec{g}_t = [dx\ dy\ dz] = \left[\frac{\partial I_j(p)}{\partial x}\ \frac{\partial I_j(p)}{\partial y}\ \frac{\partial I_j(p)}{\partial t} \right], \tag{2.12}$$

ou em coordenadas esféricas:

$$\vec{s}_t = [\rho_p\ \theta_p\ \psi_p], \tag{2.13}$$

com $\theta \in [0, \pi]$, $\psi \in [0, 2\pi)$ e $\rho = ||\vec{g}_t||$. Esse vetor indica a direção de maior variação de brilho que pode ser resultado de movimento local.

O gradiente dos n pontos de uma imagem I_j pode ser representado por um histograma

tridimensional de gradientes $\vec{h}_j = \{h_{l,k}\}$, $k \in [1, b_\theta]$ e $l \in [1, b_\psi]$, onde b_θ e b_ψ são o número de células para as coordenadas θ e ψ respectivamente. O histograma é calculado da seguinte forma:

$$h_{l,k} = \sum_p \rho_p, \tag{2.14}$$

onde $\{p \in I_j | k = 1 + \lfloor \frac{b_\theta \cdot \theta_p}{\pi} \rfloor, l = 1 + \lfloor \frac{b_\psi \cdot \psi_p}{2\pi} \rfloor\}$ são todos os pontos cujos ângulos são mapeados no intervalo da célula (k, l). O gradiente é então representado por um vetor de $b_\theta \cdot b_\psi$ elementos.

Para adicionar uma maior correlação espacial e aumentar a taxa de reconhecimento, cada quadro do vídeo é particionado em subjanelas e é calculado um histograma de gradientes para cada uma delas em separado. Assim, cada quadro é dividido em $n_x \times n_y$ partições não sobrepostas e para cada partição é calculado um histograma $\vec{h}_j^{a,b}$, $a \in [1, n_x]$ e $b \in [1, n_y]$. Pode-se ainda fazer uma reflexão horizontal do quadro a fim de reforçar simetrias horizontais do gradiente.

Na Figura 2.8, é apresentado um exemplo do cálculo do HOG. Na primeira etapa é calculada a magnitude e a orientação do gradiente para cada ponto na região em torno do ponto chave, utilizando a sua escala para selecionar o nível de suavização da gaussiana. Para obter invariância relativamente à orientação, as coordenadas do descritor e as orientações do gradiente são rodadas relativamente à orientação do ponto chave. Na fase seguinte é utilizada uma função de peso gaussiana com σ igual a metade da largura da janela para atribuir o peso à magnitude de cada ponto. O objetivo da utilização desta função é evitar alterações bruscas no descritor com pequenas variações na posição da janela e dar uma menor relevância aos pontos mais distantes do centro.

Do lado direito da Figura 2.8 podemos ver o descritor. Consiste numa sub-região de 4×4 que acumula os gradientes em histogramas de orientação com 8 direções, em que o valor de cada uma das setas representa a magnitude do histograma nessa direção. O descritor é formado por um vetor que contém todos os valores dos histogramas, correspondentes ao tamanho de cada uma das setas. No exemplo dado, os histogramas orientados formam um vetor de dimensão 2×2, este tamanho pode ser variável. O tamanho n da região de $n \times n$ dos histogramas orientados e o número de direções d a calcular são os parâmetros utilizados para variar a complexidade do descritor sendo o seu tamanho igual a $d \cdot n^2$. Segundo

Lowe (2004), quanto maior for o tamanho do descritor, maior será a sua capacidade de diferenciar em grandes conjuntos sendo, no entanto, mais propício a distorções na forma e a oclusões.

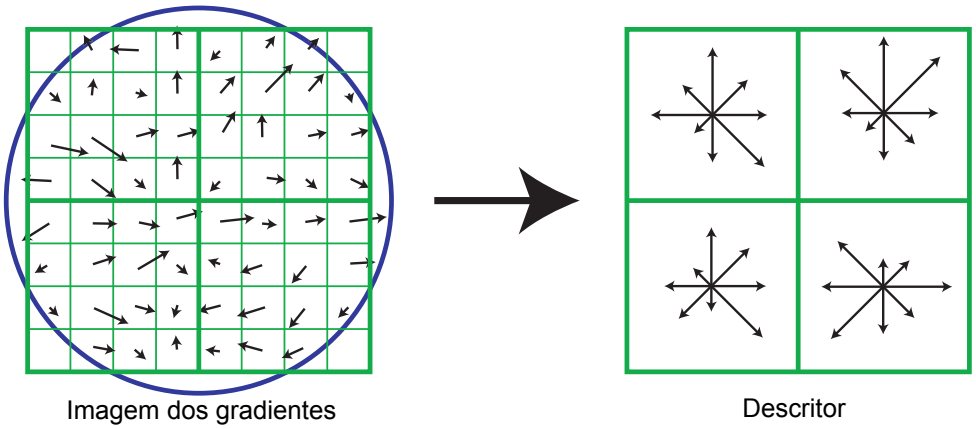


Figura 2.8: Exemplo do cálculo do descritor HOG (LOWE, 2004).

2.5 MÁQUINA VETOR SUPORTE

Tendo como base a Teoria da Aprendizagem Estatística, a Máquina de Vetores Suporte (SVM), foi desenvolvida por Vapnik (VAPNIK, 1995), com o intuito de resolver problemas de classificação de padrões. Segundo Haykin (HAYKIN, 2001) a máquina de vetores suporte é uma outra categoria das redes neurais alimentadas adiante, ou seja, redes cujas saídas dos neurônios de uma camada alimentam os neurônios da camada posterior, não ocorrendo a realimentação. Esta técnica originalmente desenvolvida para classificação binária, busca a construção de um hiperplano como superfície de decisão, de tal forma que a separação entre as classes seja máxima, considerando classes linearmente separáveis. Para classes não linearmente separáveis, busca-se uma função de mapeamento apropriada para conseguir aumentar a dimensionalidade a fim de tornar o conjunto mapeado linearmente separável. Devido a sua eficiência em trabalhar com dados de alta dimensionalidade, é reportada na literatura como uma técnica altamente robusta, muitas vezes comparada as Redes Neurais (SUNG; MUKKAMALA, 2003).

2.5.1 CLASSES LINEARMENTE SEPARÁVEIS

Uma classificação linear consiste em determinar uma função $f : X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^n$ que atribui um rótulo $(+1)$ se $f(x) > 0$ e (-1) caso contrário. Considerando uma função

linear, podemos representá-la pela Equação 2.16:

$$f(x) = \langle w \cdot x \rangle + b \tag{2.15}$$

$$= \sum_{i=1}^{n} w_i x_i + b \tag{2.16}$$

onde w e $b \in \mathbb{R}^n \times \mathbb{R}^n$, são conhecidos como vetor peso e *bias*, sendo estes parâmetros responsáveis por controlar a função e a regra de decisão. Os valores de w e b são obtidos pelo processo de aprendizagem a partir dos dados de entrada.

O vetor peso (w) e o *bias* (b) podem ser interpretados geometricamente sobre um hiperplano. Um hiperplano é um subespaço afim, que divide um espaço em duas partes, correspondendo a dados de duas classes distintas.

Sendo assim um SVM linear busca encontrar um hiperplano que separe perfeitamente os dados de cada classe e cuja margem de separação seja máxima, sendo denominado de hiperplano ótimo (Fig. 2.9).

Esse hiperplano ótimo pode ser definido matematicamente como:

$$f(x) = \langle w \cdot x \rangle + b = 0 \tag{2.17}$$

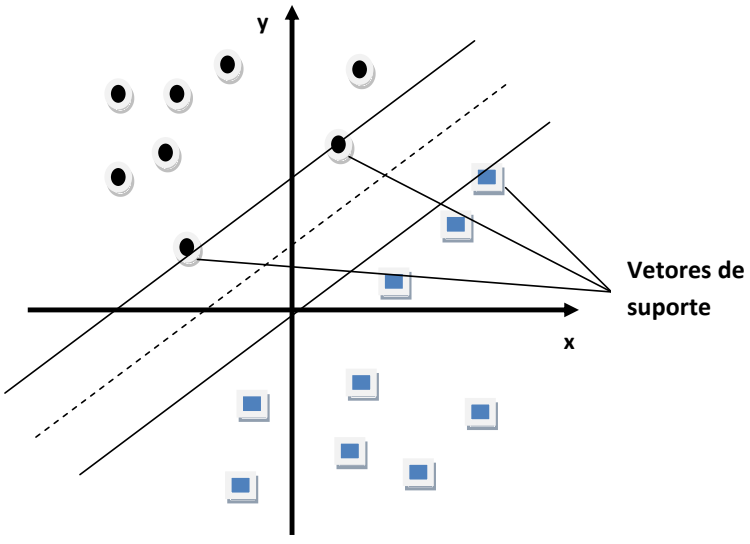


Figura 2.9: Exemplo de duas classes separadas por um hiperplano ótimo.

2.5.2 CLASSES NÃO LINEARMENTE SEPARÁVEIS

Em problemas reais dificilmente será encontrado um caso onde os dados serão linearmente separáveis, a maioria dos problemas atuais são complexos e não-lineares. Para estender a SVM linear para resolução de problemas não lineares, foram introduzidas funções reais, que mapeiam o conjunto de treinamento em um espaço linearmente separável, o espaço de características.

Um conjunto de dados é dito ser não linearmente separável, caso não seja possível separar os dados com um hiperplano.

O teorema de Cover afirma que um problema não-linear tem maior probabilidade de ser linearmente separável, em um espaço de mais alta dimensionalidade. A partir disso, a SVM não-linear realiza uma mudança de dimensionalidade, por meio das funções *Kernel*, caindo então em um problema de classificação linear, podendo fazer uso do hiperplano ótimo (SMOLA; BARTLETT, 2000)(Fig. 2.10).

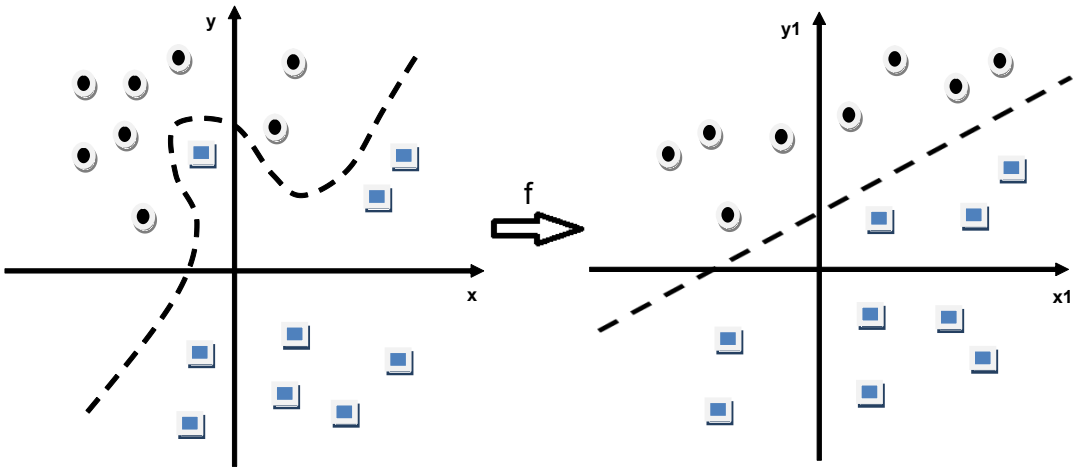


Figura 2.10: Os vetores são levados a uma dimensão maior por meio de uma função kernel f para que seja possível encontrar um hiperplano separador.

3 DESCRITOR TENSORIAL PROPOSTO

Nesta dissertação, assume-se que movimento pode ser detectado através da aplicação de filtros passa-banda em cada quadro de um vídeo. Um vídeo $\mathbf{V}$ é definido como uma sequência de quadros $\{\mathbf{I}_1,\mathbf{I}_2,\cdots, \mathbf{I}_n\}$, com n sendo o número de imagens e $\mathbf{I}_i \subset \mathbb{R}^2$.

A aplicação de múltiplos filtros é usada para extrair diferentes espectros do vídeo original. O ponto chave deste trabalho é que cada filtro correlaciona o espectro original de maneira distinta, e isso é usado para capturar nuâncias do movimento. A motivação para isso reside no fato de que apenas um simples operador como o Sobel, aplicado depois de um operador gaussiano, pode conseguir $92,1\%$ (PEREZ et al., 2012) de taxa de reconhecimento na base KTH. A informação de movimento extraído do vídeo $\mathbf{V}$ é representada de forma compacta através do uso de histogramas de gradiente (Sec. 2.2.2.2) e tensores de orientação (Sec. 2.2.2.2).

3.1 EXTRAÇÃO DE MOVIMENTO COM MÚLTIPLOS FILTROS DERIVATIVOS

Como visto em Perez (2012), ruído é um dos fatores que diminuem a capacidade de extrair movimento em um vídeo. O primeiro passo para extração de movimento no vídeo $\mathbf{V}$ consiste na convolução de um filtro gaussiano B em cada quadro $\mathbf{I} \in \mathbf{V}$. A resposta de impulso da gaussiana é mostrada na Figura 3.1.

0.006	0.061	0.242	0.383	0.242	0.061	0.006

Figura 3.1: Máscara gaussiana unidimensional.

Na sequência do processamento, definimos **V'**, resultado da convolução da máscara gaussiana B na direção x e y separadamente, como uma sequência de quadros $\{\mathbf{Q}_1, \mathbf{Q}_2, \cdots, \mathbf{Q}_n\} \mid \mathbf{Q}_k = (B * \mathbf{I}_k)$, com n sendo o número de imagens e $\mathbf{I} \in \mathbb{R}^2$. Essa filtragem serve para atenuar as altas frequências, que podem representar algum tipo de ruído que não seja movimento. É importante ressaltar que todos os procedimentos a seguir são baseados no novo vídeo produzido **V'**.

3.1.1 FILTROS DERIVATIVOS

Podemos definir um filtro derivativo unidimensional por um par de respostas de impulso (H_a, G_a), onde $a \in \{1, 2, \cdots, f\}$ é o índice do filtro, f é o número de filtros disponíveis para realizar a detecção de movimento, G_a tem a resposta de frequência de um passa-alta, e H_a tem a resposta de frequência de um passa-baixa. A versão multidimensional dos filtros são separáveis, tendo H_a e G_a como fatores. Devido à aplicação do filtro gaussiano B em cada quadro do vídeo $\mathbf{V}$, o impulso de resposta do filtro G_a sofre uma substancial modificação, já que determinadas altas frequências contidas no vídeo original $\mathbf{V}$ foram atenuadas, ou mesmo eliminadas, durante a produção do novo vídeo $\mathbf{V'}$ (Fig. 3.2).

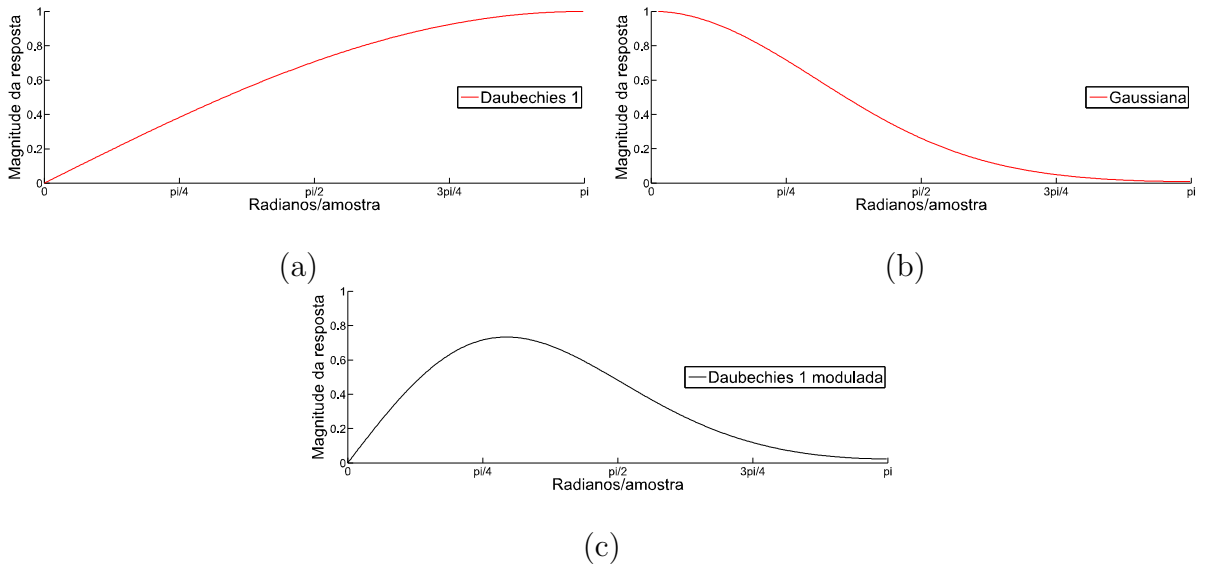


Figura 3.2: Função de transferência do filtro Daubechies 1 modulado pelo filtro Gaussiano B nos eixos $\boldsymbol{x}$ e $\boldsymbol{y}$. (a) Função de transferência do filtro Daubechies 1. (b) Função de transferência do filtro gaussiano. (c) Função final de transferência da convolução $(B * G_{db1})$, onde G_{db1} representa o filtro passa-alta Daubechies 1.

Os filtros derivativos são usados para capturar a informação de movimento contida em uma sequência de quadros $\mathbf{Q}_k$ do vídeo $\mathbf{V'}$. Desta forma, a resposta de impulso G_a é usado como um estimador de gradiente com resposta de frequência $\tilde{G_a}$. Para sinais multidimensionais, H_a atenua o ruído nas direções ortogonais. As abordagens baseadas em gradiente fornecem uma estimativa do movimento através das variações de brilho ocorridas em cada imagem. Estas mudanças são modeladas por médias de equações diferenciais parciais, que são geralmente chamadas como equações de restrição.

Neste trabalho, assume-se que a resposta de frequência $\tilde{H_a}$ possui um grau de complementaridade em relação a $\tilde{G_a}$, a fim de atenuar o ruído correlacionado indesejado entre os

eixos principais.

As derivadas parciais, ou gradiente, resultado da aplicação de um filtro (H_a, G_a) sobre o k-ésimo quadro $\mathbf{Q}_k$ do vídeo **V'**, no ponto p, é definida de acordo com:

$$\vec{g} = [dx_p^a \; dy_p^a \; dt_p^a]^T = \left[\frac{\partial Q_k(p)}{\partial x} \; \frac{\partial Q_k(p)}{\partial y} \; \frac{\partial Q_j(p)}{\partial t}\right]^T. \tag{3.1}$$

O componente dx_p^a é calculado pela convolução do a-ésimo filtro no vídeo **V'** em relação aos eixos $\boldsymbol{x,y,t}$ da seguinte forma:

- convolução do filtro H_a em relação ao eixo $\boldsymbol{y}$;
- convolução do filtro H_a em relação ao eixo $\boldsymbol{t}$;
- convolução do filtro G_a em relação ao eixo $\boldsymbol{x}$.

É importante observar que dx_p^a é calculado sobre o vídeo **V'**, portanto, sofre a influência do filtro gaussiano B.

O cálculo do componente dy_p^a ocorre da seguinte forma:

- convolução do filtro H_a em relação ao eixo $\boldsymbol{x}$;
- convolução do filtro H_a em relação ao eixo $\boldsymbol{t}$;
- convolução do filtro G_a em relação ao eixo $\boldsymbol{y}$.

Da mesma forma que ocorre com dx_p^a, o componente dy_p^a sofre influência do filtro gaussiano B.

Por fim, para calcular o componente dt_p^A devemos prosseguir da seguinte maneira:

- convolução do filtro H_a em relação ao eixo $\boldsymbol{x}$;
- convolução do filtro H_a em relação ao eixo $\boldsymbol{y}$;
- convolução do filtro G_a em relação ao eixo $\boldsymbol{t}$.

Em relação à convolução realizada no eixo $\boldsymbol{t}$, deve-se ressaltar que cada ponto ao longo deste eixo representa um quadro $\mathbf{Q}_k$ do vídeo **V'**. Portanto, uma convolução realizada neste eixo leva em consideração uma determinada quantidade de quadros $\{\mathbf{Q}_1, \mathbf{Q}_2, \cdots, \mathbf{Q}_n\}$ do vídeo **V'**, onde n é definido pelo número de coeficientes do filtro escolhido para ser utilizado. Note que dt_p^a também sofre a influência do filtro gaussiano B apenas nas direções ortogonais $\boldsymbol{x}$ e $\boldsymbol{y}$.

3.1.2 FILTRO DE CORRELAÇÃO

O espectro de um vídeo **V** é determinado pelo filtro derivativo (H_a, G_a), onde a representa o índice de um determinado filtro selecionado, aplicado sobre cada um dos quadros $\mathbf{Q}_n$ que o compõe. Por isso, pode-se afirmar que cada filtro aplicado sobre um determinado vídeo nos permite realizar uma análise específica de algum tipo de fenômeno ocorrido em sua sequência de quadros. Com intuito de extrair diferentes espectros de um mesmo vídeo, é possível combinar a resposta obtida pela aplicação de vários filtros.

Para correlacionar os filtros, é proposto a derivação de um filtro (H_{f+1}, G_{f+1}) tal que:

$$|\widetilde{H}_{f+1}(\omega)| = \sum_{a=1}^{f} |\widetilde{H}_a(\omega)|$$

,

$$|\widetilde{G}_{f+1}(\omega)| = \sum_{a=1}^{f} |\widetilde{G}_a(\omega)|,$$

ou seja, a magnitude da resposta é a mesma que a soma das magnitudes dos $f > 1$ filtros.

Com o filtro projetado para correlacionar múltiplos espectros é possível melhorar a análise de movimento de um vídeo.

3.2 COMPUTANDO HOG3D EM CADA QUADRO

A saída filtrada de um quadro $\mathbf{Q}_k$, com n pontos p, pode ser compactamente representada por um histograma tridimensional de gradientes $\vec{h}_k^a = \{h_{j,l}^a\}$, $j \in [1, nb_\theta]$ e $l \in [1, nb_\psi]$, onde nb_θ e nb_ψ são o número de células para as coordenadas θ e ψ respectivamente. Existem vários métodos para calcular o HOG3D e escolhemos, pela sua simplicidade, uma subdivisão uniforme do intervalo de ângulos para preencher as $nb_\theta \cdot nb_\psi$ classes:

$$h_{j,l}^a = \sum_p \rho_p^a \cdot w(dist_{j,l}^{q,r}),$$

onde $dist_{j,l}^{q,r}$ é a distância euclidiana entre a classe de índice (j, l) e o mapeamento das coodernadas reais $(q, r) = (1 + \frac{nb_\theta \cdot \theta_p^a}{\pi}, 1 + \frac{nb_\psi \cdot \psi_p^a}{2\pi})$ do gradiente no ponto p, e $w(dist_{j,l}^{q,r})$ é uma função de ponderação gaussiana com $\alpha = 1,0$ (LOWE, 1999). O gradiente do k-ésimo quadro **Q** do vídeo **V'** é então representado por um vetor $\vec{h}_k^a$ com $nb_\theta \cdot nb_\psi$ elementos. Todos os resultados produzidos nesta dissertação são computados usando $nb_\theta = 8$ e $nb_\psi = 16$

(PEREZ et al., 2012). Vale ressaltar que o HOG3D é calculado em todos os quadros $\mathbf{Q}_k$ do vídeo **V'** para cada filtro (H_a, G_a) escolhido.

Para reduzir a diferença de brilho entre cada quadro do vídeo, o histograma de gradientes $\vec{h}_k^a \in \mathbb{R}^{nb_\theta \cdot nb_\psi}$ pode ter todos seus elementos $h_{j,l}^a$ ajustados para ${h_{j,l}^a}^\gamma$, com $\gamma = 0,5$. Esse processo é chamado de normalização de energia (*power normalization*) e serve para reduzir a diferença entre as classes do gradiente. Esta técnica é aplicada somente em alguns resultados, com intuito de melhorar o desempenho dos descritores.

3.3 TENSOR DE ORIENTAÇÃO: CODIFICANDO COEFICIENTES DO HOG3D

Um tensor de orientação, como visto na Seção 2.2.2.2, é uma matriz $m \times m$ real e simétrica, para sinais m-dimensionais. É importante notar que um tensor de estrutura bem conhecido é um caso específico de um tensor de orientação (JOHANSSON et al., 2002). O tensor do quadro $\mathbf{Q}_k$ usando o filtro de índice a é:

$$T_k^a = \vec{h}_k^a \vec{h_k^a}^T,$$

que carrega a informação da distribuição do gradiente do k-ésimo quadro, calculado usando o a-ésimo filtro. Individualmente, este tensor tem a mesma informação de $\vec{h}_k^a$. Uma vez que T_k^a é uma matriz simétrica, ele pode ser armazenado com $\frac{m(m+1)}{2}$ elementos.

Para um filtro derivativo de índice a, temos que expressar a média de movimento dos quadros consecutivos utilizando uma série de tensores. O movimento médio de um vídeo pode ser determinado por:

$$T^a = \sum_{k=1}^{n} T_k^a \tag{3.2}$$

onde n é o número de quadros do vídeo. Pode ser usado todos os quadros do vídeo ou apenas um intervalo de interesse. Normalizando T^a com uma norma l_2 , nos permite realizar uma comparação entre vídeos, independentemente do seu comprimento ou resolução da imagem.

Se a série de acumulação diverge, obtém-se um tensor isotrópico que não contém informações úteis extraídas pelo par de estimadores derivativos de índice a. Porém, se a série convergir, tem-se um tensor anisotrópico que transporta a informação de movimento

mais significativo da sequência de quadros analisados.

3.3.0.1 Subdivisão dos quadros

Quando um histograma de gradiente é calculado usando a imagem inteira, suas células são preenchidas com vetores, independentemente da sua posição na imagem. Isto implica em uma perda de correlação entre os vetores de gradiente e seus vizinhos. Como observado em vários trabalhos (LOWE, 1999), a subdivisão do vídeo em cubos proporciona uma melhor taxa de reconhecimento (Fig. 3.3).

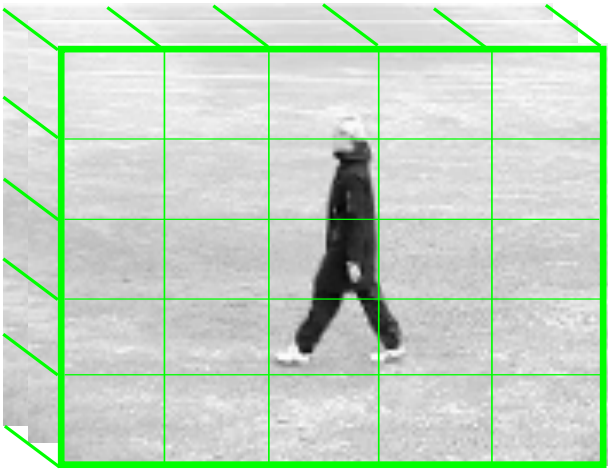


Figura 3.3: Subdivisão do vídeo em cubos.

Supondo que o quadro $\mathbf{Q}_k$ do vídeo **V'**, seja uniformemente subdividido nas direções $\boldsymbol{x}$ e $\boldsymbol{y}$ formando uma grade com n_x e n_y blocos não sobrepostos. Cada bloco pode ser visto como um vídeo distinto variando no tempo. As subimagens resultam no histograma de gradiente $\vec{h}_k^a(c,r)$, $c \in [1, n_x]$ e $r \in [1, n_y]$, em que os vetores de gradiente possuem uma melhor correlação local entre si. O tensor para o quadro $\mathbf{Q}_k$, usando o a-ésimo filtro derivativo, é então calculado como a soma dos tensores de cada bloco:

$$T_k^a(c,r) = \sum_{c,r} \vec{h}_k^a(c,r)\ \vec{h}_k^a(c,r)^{\,T},$$

capturando a incerteza da direção do histograma de vetores m-dimensionais $\vec{h}_k^a(c,r)$ para o quadro $\mathbf{Q}_k$. A série de tensores torna-se:

$$T^a = \sum_{k=1}^{n} \sum_{c=1}^{n_x} \sum_{r=1}^{n_y} \frac{T_k^a(c,r)}{||T_k^a(c,r)||},$$

onde a é o índice do filtro derivativo usado, k é o índice do quadro do vídeo **V'**, e $(c,r) \in$

$[1, n_x] \times [1, n_y]$ são as coordenadas das subimagens.

O descritor tensorial final do vídeo **V'** para o filtro derivativo a é dado por $\frac{T^a}{||T^a||}$, esse descritor contém o mesmo número de elementos da versão sem subdivisão da imagem.

3.4 DESCRITOR TENSORIAL GLOBAL: CONCATENANDO TENSORES BASEADOS EM MÚLTIPLOS FILTROS

Os descritores de vídeos podem ser classificados de duas maneiras:

Descritores locais: que focam em determinados pontos de uma imagem, tentando extrair algum tipo de característica especial. O método conhecido como *Scale-invariant feature transform - SIFT* (LOWE, 1999), é um exemplo de descritor local que faz uma busca na imagem procurando por pontos de interesse que apresentam invariância em relação à posição, escala e localização.

Descritores globais: que visam descrever todo o contéudo do vídeo. A principal vantagem do uso de descritores globais é sua simplicidade, já que não há necessidade de um conhecimento prévio do vídeo a ser analisado (MOTA, 2011). Podemos definir um descritor global de movimento como um par - vetor de características extraídas e função de distância - usado para indexação por similaridade de vídeos e/ou imagens. O vetor de características contém as propriedades da imagem ou do vídeo e a função de distância mede a similaridade entre duas imagens ou dois vídeos. Na maioria das vezes, a similaridade é definida como inversa à função de distância (por exemplo, distância Euclidiana), assim, quanto menor a distância entre as imagens ou vídeos, maior é a similaridade entre eles.

O ponto chave desta dissertação é usar uma correlação entre os tensores, calculados para todos os pares de filtro (H_a, G_a) onde $a \in \{1, 2, \cdots, f\}$, a fim de conseguir melhores resultados para o reconhecimento de ações humanas em vídeos. Uma maneira de combiná-los é através da concatenação desses tensores. Portanto, o descritor tensorial final $\mathbf{T}$ para o vídeo de entrada $\mathbf{V}$ é dado por:

$$\mathbf{T} = \{T^1, T^2, \cdots, T^a\}.$$

Apesar de outros métodos de combinação serem possíveis, a concatenação entre os

descritores preserva a informação de movimento extraído por cada filtro. A desvantagem é que o número de coeficientes no descritor é multiplicado pelo número de filtros derivativos utilizados. Neste trabalho, o HOG3D tem 128 classes produzindo tensores com 8256 elementos para um único filtro. Um descritor de vídeo utilizando quatro filtros, por exemplo, têm 33024 elementos, tornando a classificação pelo SVM mais custosa.

4 RESULTADOS

Neste capítulo, apresenta-se os resultados obtidos com o descritor de movimentos proposto, comparando-o aos resultados mais recentes encontrados na literatura. Para validar nosso descritor usamos a base de dados KTH.

O protocolo de classificação utilizado foi baseado na estratégia conhecida como *two-fold* (SOLMAZ et al., 2012) com um classificador SVM não linear de *kernel* gaussiano. Todos os resultados foram computados usando $nb_\theta = 8$ e $nb_\psi = 16$, tendo um HOG3D com 128 classes por quadro do vídeo (PEREZ et al., 2012). O tensor de um filtro possui então 8256 elementos.

A classificação dos descritores foi realizada no sistema RETIN (*REcherche et Traque INteractive d'images)* do laboratório ETIS (*Equipes Traitement de l'Information et Systèmes*) da ENSEA (*École Nationale Supérieure de l'Électronique et de ses Applications*) (FOURNIER et al., 2001).

4.1 BASE DE DADOS KTH

A base de vídeos KTH é composta por 6 tipos de movimentos:

1. *Walking*: movimento de pessoa caminhando;
2. *Jogging*: movimento entre uma corrida e uma caminhada;
3. *Running*: movimento de pessoa correndo;
4. *Boxing*: movimento de pessoa desferindo socos no ar;
5. *Hand waving*: movimento de pessoa agitando os braços;
6. *Hand clapping*: movimento de pessoa batendo palmas.

Para um melhor entendimento dos resultados, os movimentos oriundos da base KTH recebem as seguintes abreviações: *walking* passa a ser chamado de *Walk*, *jogging* passa a ser *Jog*, *running* passa a ser *Run*, *boxing* passa a ser *Box*, *hand waving* passa a ser *HWav* e *hand clapping* passa a ser *HClap*.

Todos os resultados da base KTH foram obtidos através da classificação de cada um dos 2391 vídeos contidos nesta base.

4.2 FILTROS UTILIZADOS

Nesta seção, são mostrados os principais filtros utilizados neste trabalho. Entre eles, destacam-se os filtros Daubechies (dbn), onde n é o índice do filtro. Os gráficos da resposta de impulso dos principais filtros são mostrados nas Figuras 4.1, 4.2, 4.3 e 4.4. Vale ressaltar que como o vídeo original $\mathbf{V}$ sofre uma convolução do filtro gaussiano B em cada quadro $\mathbf{Q}_k$, a função de transferência de cada um dos filtros derivativos é substancialmente modificada. Com isso, o estudo dos filtros é baseado em sua resposta de impulso modulada pelo filtro gaussiano B. Optou-se por usar filtros wavelets como estimadores derivativos pois seu comportamento é bem conhecido. Todas as respostas de fase do filtros são omitidos, pois em todos os casos essa resposta é linear.

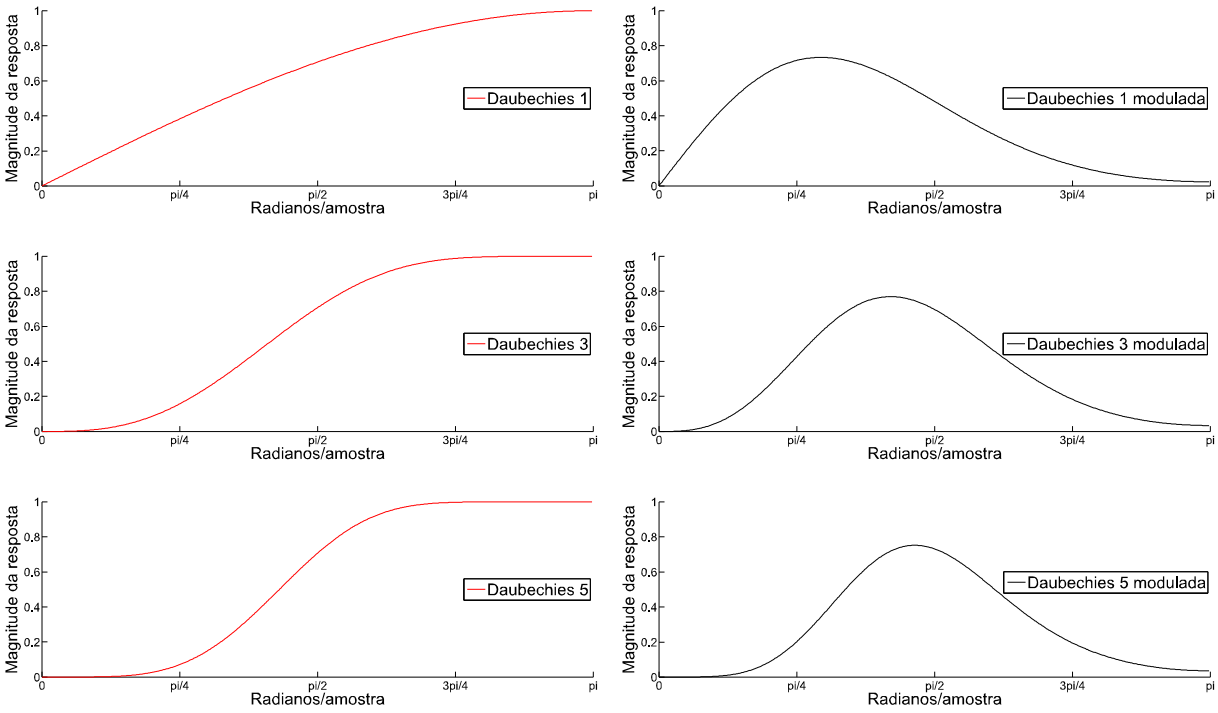


Figura 4.1: Função de transferência dos filtros $db1$, $db3$ e $db5$, modulados pelo filtro Gaussiano B nos eixos $\boldsymbol{x}$ e $\boldsymbol{y}$.

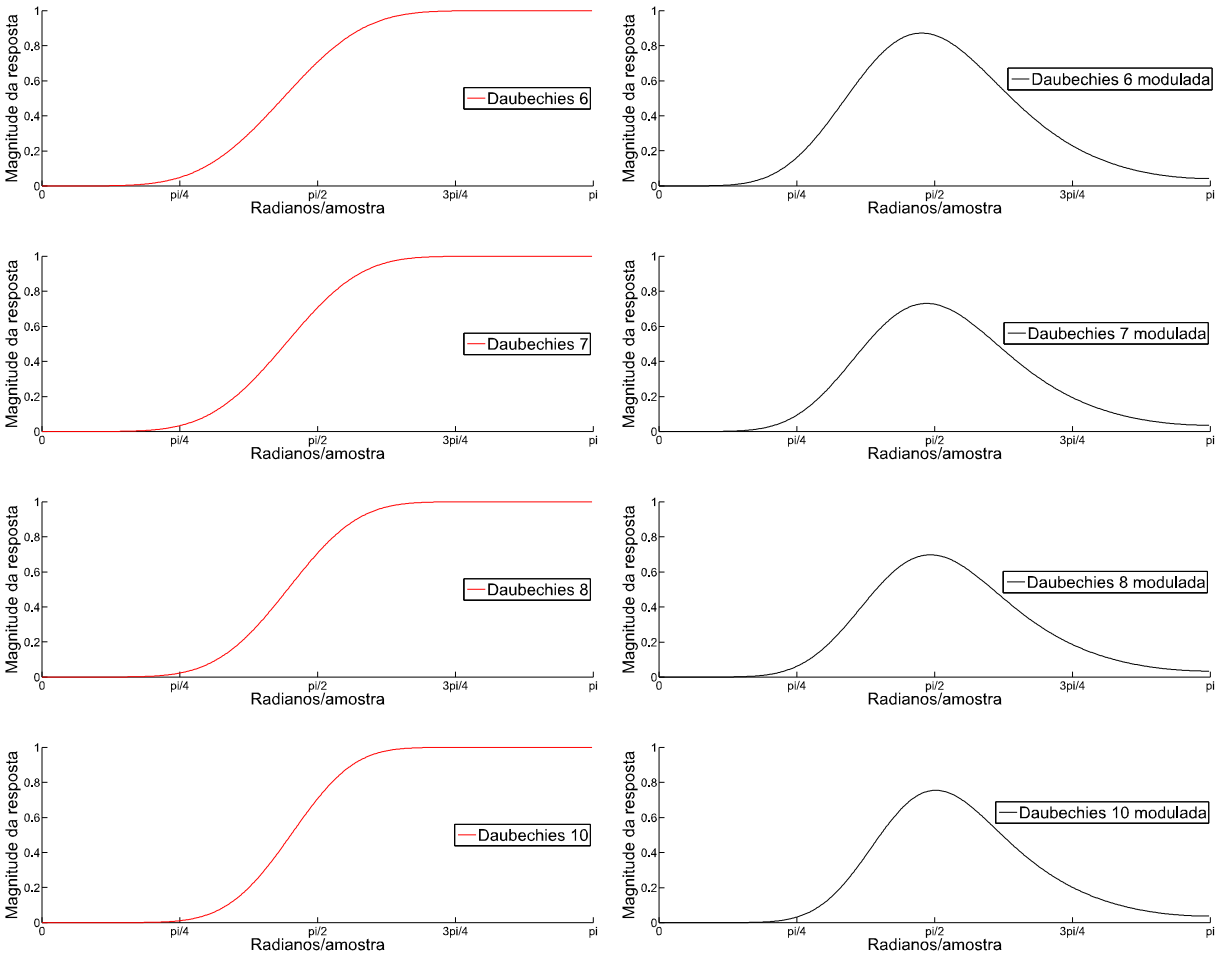


Figura 4.2: Função de transferência dos filtros *db*6, *db*7, *db*8 e *db*10, modulados pelo filtro Gaussiano B nos eixos $\boldsymbol{x}$ e $\boldsymbol{y}$.

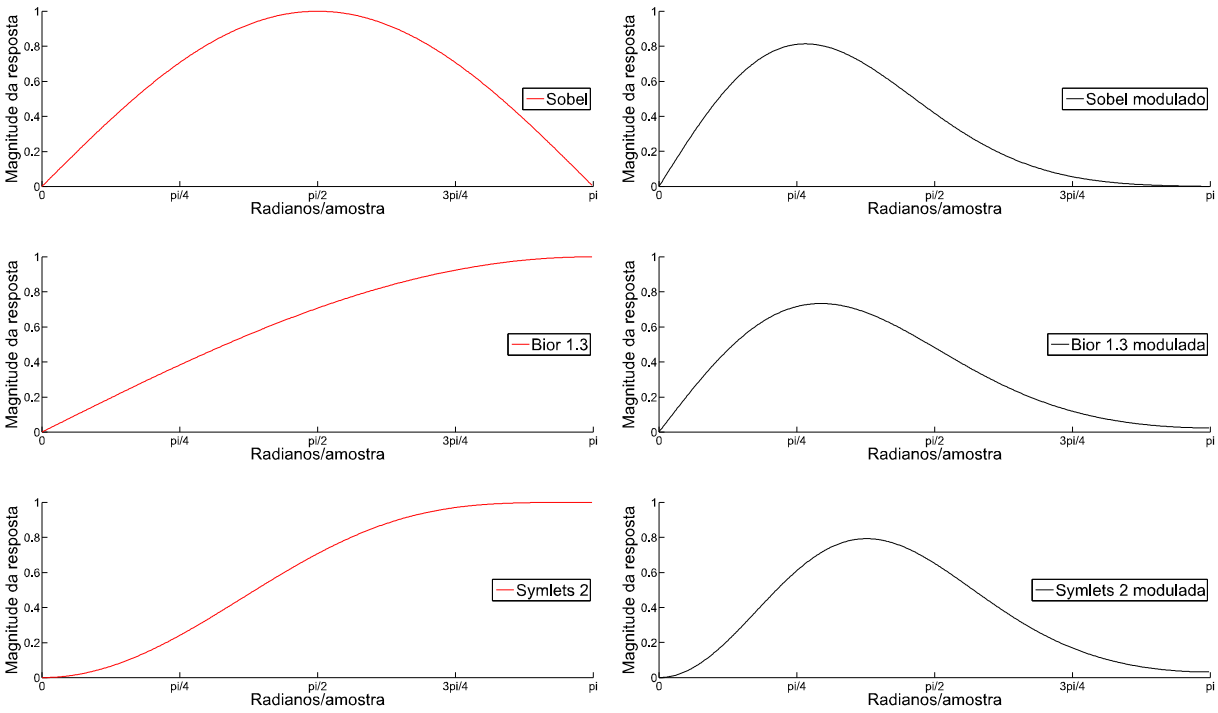


Figura 4.3: Função de transferência dos filtros *sobel*, *bior*1.3, *sym*2, modulados pelo filtro Gaussiano B nos eixos $\boldsymbol{x}$ e $\boldsymbol{y}$.

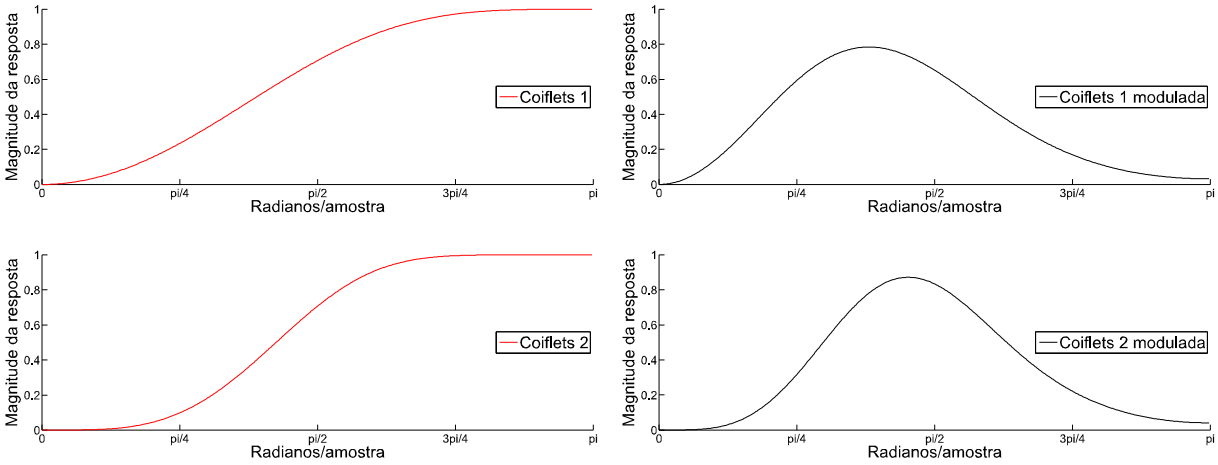


Figura 4.4: Função de transferência dos filtros $coif1$, $coif2$, modulados pelo filtro Gaussiano B nos eixos $\boldsymbol{x}$ e $\boldsymbol{y}$.

Todos os filtros derivativos utilizados neste trabalho são do tipo *FIR*, logo esses filtros não são recursíveis, apresentando uma boa estabilidade (Seção 2.1.2).

O filtro *db*1, também caracterizado como filtro wavelet de Haar, não possui uma boa frequência de corte, já que não consegue separar as altas e baixas frequências de maneira satisfatória. Observa-se que a resposta de impulso do filtro *db*1 modulado por uma gaussiana B, preserva melhor as baixas frequências no primeiro quarto do espectro, se comparadas aos filtros *db*3 e *db*7 (Fig. 4.5).

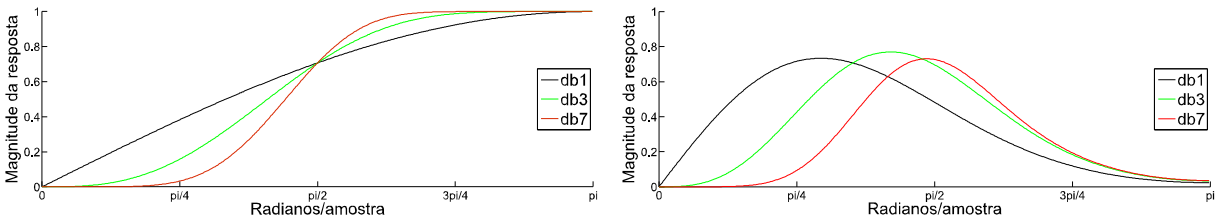


Figura 4.5: Função de transferência dos filtros *db*1, *db*3 e *db*7 modulados pela gaussiana B.

Comparando *db*2, *db*4 e *db*5, fica evidente que a frequência de corte tende a $\pi/2$ à medida que a quantidade de momentos nulos aumenta em cada filtro. Por isso, pode-se dizer que o filtro *db*5 possui uma frequência de corte um pouco mais refinada se comparado aos filtros *db*2 e *db*4. A Figura 4.6 mostra que a resposta de impulso do filtro *db*5 modulado pela gaussiana B, não consegue preservar as baixas frequências da mesma forma que os filtros *db*2 e *db*4, porém, preserva mais altas frequências no terceiro quarto do espectro do que os outros filtros.

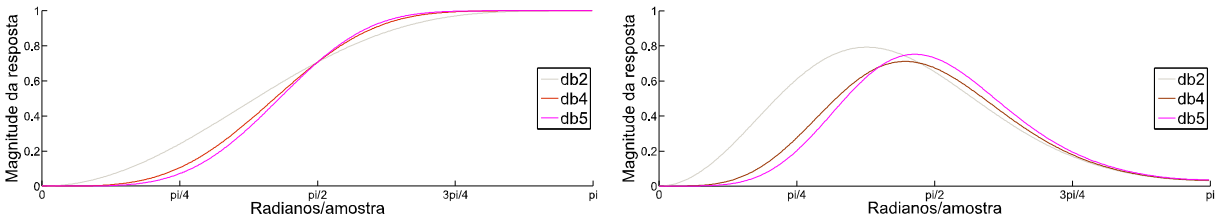


Figura 4.6: Função de transferência dos filtros *db*2, *db*4 e *db*5 modulados pela gaussiana B.

O estudo comparativo realizado para o filtro *db*8, apresenta uma resposta de impulso modulado pela gaussiana B, praticamente centrada no eixo do espectro. Pode-se dizer então, que o filtro não consegue preservar a mesma quantidade de baixas frequências como o *db*6.

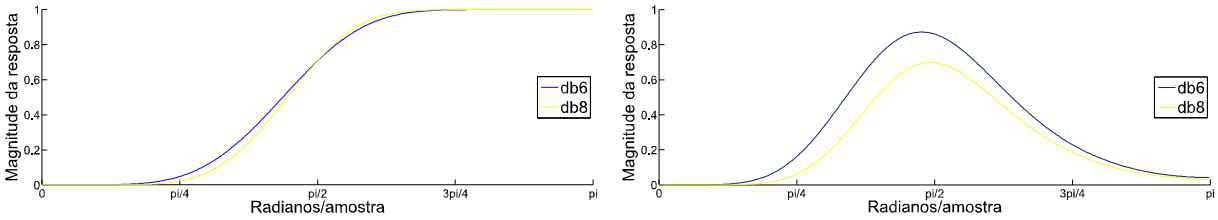


Figura 4.7: Função de transferência dos filtros *db*6 e *db*8 modulados pela gaussiana B.

Os filtros *db*9 e *db*10, por apresentarem uma maior quantidade de momentos nulos, possuem uma boa frequência de corte no espectro, pois conseguem separar as baixas das altas frequências. A resposta de impulso de ambos os filtros é bem parecida, o que as diferencia é o fato do filtro *db*10 conseguir preservar um pouco mais das altas frequências que o filtro *db*8 e *db*9(Fig. 4.8).

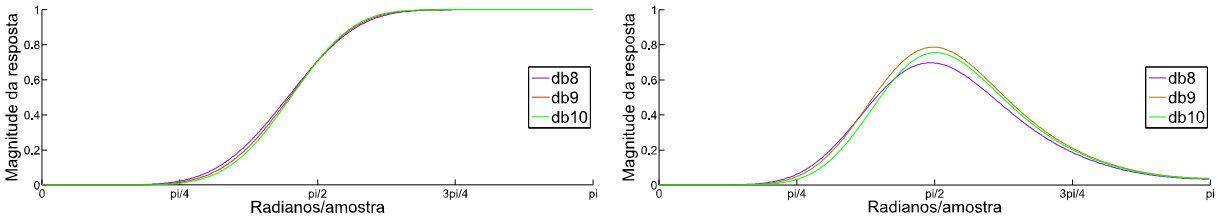


Figura 4.8: Função de transferência dos filtros *db*8, *db*9 e *db*10 modulados pela gaussiana B.

4.3 SUBDIVISÃO DOS QUADROS

Através da classificação da base de dados KTH, por um SVM com protocolo *two-fold*, a Figura 4.9 mostra as diferentes taxas de reconhecimento variando o número de subdivisões dos quadros.

Constata-se que ao realizar subdivisões nos quadros dos vídeos, obtém-se um aumento na taxa de reconhecimento. A ocorrência desse fenômeno está relacionada com a obtenção de uma melhor correlação de posição nos histogramas de gradiente (Seção 3.3.0.1). Na Tabela 4.1, verifica-se que os resultados em cada uma das subdivisões realizadas, com 4×4 e 8×8 partições, consegue-se melhores resultados para o filtro *db*1. Os experimentos a seguir foram realizados utilizando apenas quadros com 8×8 partições, pois para os demais

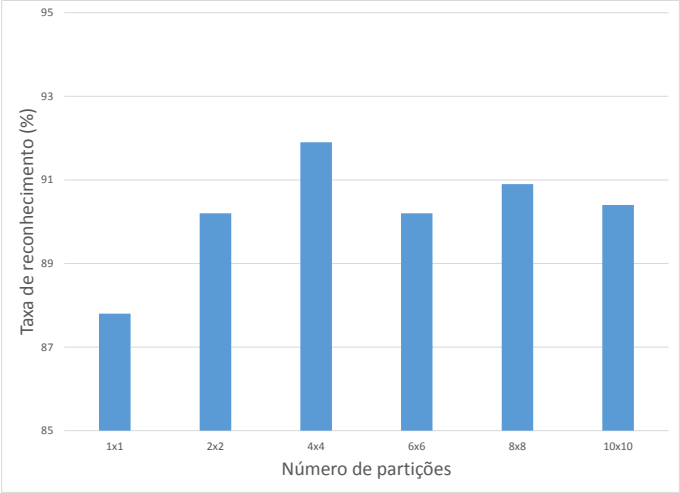


Figura 4.9: Resultado da classificação da base KTH usando filtro derivativo *db*1 com HOG 16×8.

Partições	**Taxa de reconhecimento**
1x1	87,8%
2x2	90,2%
4x4	91,9%
6x6	90,2%
8x8	90,9%
10x10	90,4%

Tabela 4.1: Taxa de reconhecimento com variação no número de subdivisões dos quadros.

filtros, constatou-se que esse número de partições apresenta resultados mais satisfatórios.

4.4 RESULTADO COM FILTROS ISOLADOS

Nesta seção, mostram-se os resultados alcançados para cada um dos filtros derivativos usados. Vale lembrar que esses resultados foram obtidos usando o classificador SVM com protocolo *two-fold*.

Como método comparativo, para comprovar que o uso de subdivisões nos quadros melhora a taxa de reconhecimento, os resultados foram gerados para dois casos: o primeiro, usando um número de partições igual a 1×1, ou seja, é usado o quadro inteiro do vídeo; o segundo caso, com 8×8 partições de cada quadro. A Figura 4.10 mostra um comparativo entre os resultados obtidos por cada filtro, sem subdivisão da imagem.

É possível observar que o filtro *db*1 apresenta um bom resultado, se comparado aos demais filtros (Tab. 4.2).

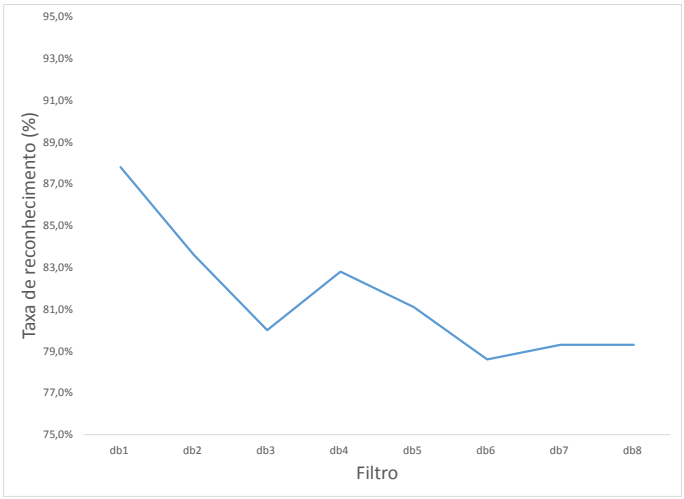


Figura 4.10: Gráfico comparativo entre os filtros sem subdivisão dos quadros.

Filtro	**Taxa de reconhecimento**
$db1$	87,8%
$bior1.3$	86,0%
$sobel$	85,2%
$coif2$	83,9%
$sym2$	83,6%
$db2$	83,6%
$coif1$	82,8%

Filtro	**Taxa de reconhecimento**
$db3$	80,0%
$db4$	82,8%
$db5$	81,1%
$db6$	78,6%
$db7$	79,3%
$db8$	79,3%

Tabela 4.2: Taxa de reconhecimento para cada filtro com partição 1×1.

A Tabela 4.3 mostra a capacidade do filtro $db1$ de capturar cada um dos movimentos ocorridos no vídeo. Vale ressaltar que este filtro consegue capturar bem os movimentos *Box*, *HWay* e *Walk*, porém, não consegue distinguir de forma satisfatória os movimentos *HClap* e *Jog*.

	Box	HClap	HWav	Jog	Run	Walk
Box	96.5	3.5	0.00	0.00	0.00	0.0
HClap	21.5	78.5	0.0	0.00	0.00	0.00
HWav	3.5	0.7	95.8	0.00	0.00	0.00
Jog	0.7	0.00	0.00	79.9	11.8	7.6
Run	0.00	0.00	0.00	17.4	80.6	2.1
Walk	0.00	0.00	0.00	2.8	1.4	95.8

Tabela 4.3: Matriz de confusão para o filtro *db*1 sem subdivisão dos quadros.

A Figura 4.11 mostra uma comparação entre os resultados obtidos para cada filtro, utilizando uma subdivisão dos quadros com 8×8 partições.

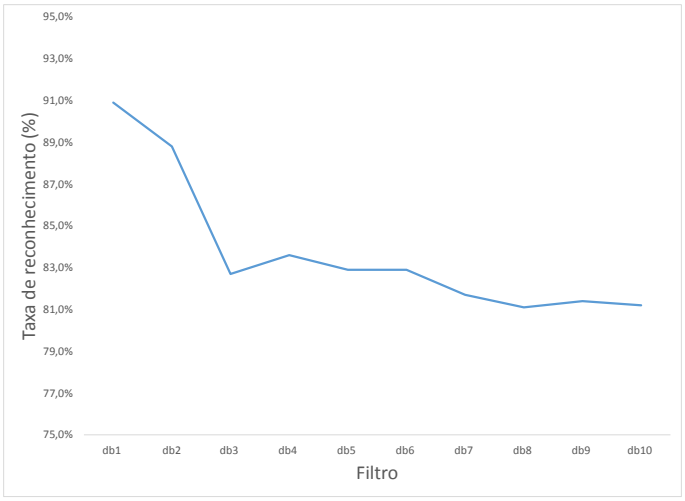


Figura 4.11: Gráfico comparativo entre os filtros com 8×8 partições.

Como demonstrado, o filtro *db*1 modulado pela gaussiana continua apresentando o melhor resultado para classificação de ações humanas em vídeos. A Tabela 4.4 mostra as taxas de reconhecimento para cada filtro. Observa-se que os resultados obtidos usando uma subdivisão dos quadros com 8×8 partições, foram superiores aos que não utilizaram esta técnica.

Filtro	Taxa de reconhecimento
$db1$	90,9%
$bior1.3$	90,6%
$sym2$	89,9%
$sobel$	88,9%
$db2$	88,8%
$coif1$	87,5%
$db4$	83,6%
$db5$	82,9%

Filtro	Taxa de reconhecimento
$db6$	82,9%
$db3$	82,7%
$Gcoif2$	82,0%
$db7$	81,7%
$db8$	81,1%
$db9$	81,4%
$db10$	81,2%

Tabela 4.4: Taxa de reconhecimento para cada filtro com 8×8 partições.

A Tabela 4.5 mostra que o filtro $db1$ consegue capturar muito bem os movimentos *Box*, *HClap*, *HWay* e *Walk*. O problema deste filtro é a dificuldade para diferenciar o movimento realizado em *Jog* e *Run*, onde é classificado erroneamente $20,8\%$ dos movimentos em *Jog* como sendo *Run*.

	Box	HClap	HWav	Jog	Run	Walk
Box	97.2	2.8	0.00	0.00	0.00	0.0
HClap	3.5	94.4	2.1	0.00	0.00	0.00
HWav	5.6	0.7	93.8	0.00	0.00	0.00
Jog	0.7	0.00	0.00	86.1	8.3	5.6
Run	0.00	0.00	0.00	20.8	77.8	1.4
Walk	0.00	0.00	0.00	3.5	0.0	96.5

Tabela 4.5: Matriz de confusão para o filtro $db1$ com 8×8 partições.

Analisando os resultados obtidos em cada filtro, é possível concluir que algumas frequências médias e altas são consideradas ruídos, enquanto algumas baixas frequências são adequadas para a classificação com o conjunto de dados KTH. Conclui-se então que os filtros que apresentaram melhores resultados, conseguem preservar melhor as baixas frequências e capturando poucas médias e altas frequências. Encontrar uma combinação adequada baseada na resposta de vários filtros, pode levar a um melhor desempenho.

4.4.1 FILTRAGEM COM EXPANSÃO DOS FILTROS

Diversos métodos utilizam wavelets como base para representação de movimento em vídeos. Com intuito de observar como os filtros wavelets respondem em escala diática,

através da compressão ou dilatação em potências de 2, são apresentados os resultados da aplicação de alguns filtros nas escalas 2 e 3 (Tab. 4.6).

Filtro	**Escala 1**	**Escala 2**	**Escala 3**
$db1$	90,9%	81,2%	73,7%
$db2$	88,8%	79,5%	73,2%
$db3$	82,7%	73,8%	66,4%

Tabela 4.6: Taxa de reconhecimento para os filtros decimados com 8×8 partições.

Neste trabalho, ao invés de realizar uma decimação na imagem e depois fazer uma convolução com o filtro derivativo, é feito uma dilatação no filtro para depois convoluir na imagem:

$$G_a^k = (G_a^{k-1}(\uparrow 2)) * H_a$$

$$H_a^k = (H_a^{k-1}(\uparrow 2)) * H_a,$$

onde k representa o fator de escala do filtro de índice a (MALLAT, 1999).

A Figura 4.12 mostra o corte no espectro do filtro $db3$ para cada escala. No nível 1 o filtro $db3$ representa o espectro com corte em π, ou seja, metade do espectro é isolado. No nível 2, 1/4 do espectro é isolado, enquanto no nível 3, é possível isolar 1/8 do espectro. Tanto no nível 2 e 3 é possível perceber que o filtro não consegue preservar altas frequências como o nível 1, por isso, algumas frequências que podem ser consideradas movimento não são capturadas, fazendo com que a taxa de reconhecimento seja inferior aos filtros no nível 1.

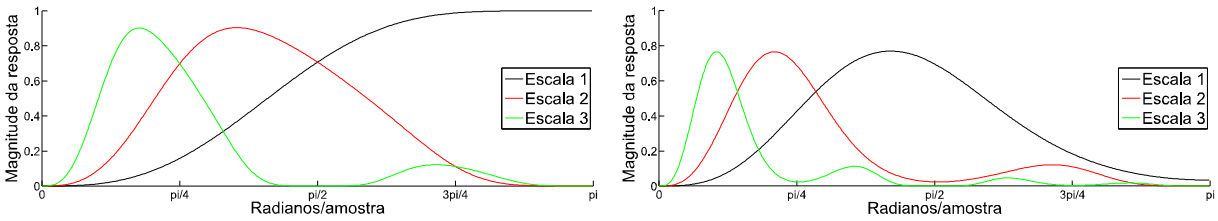


Figura 4.12: Função de transferência do filtro $db3$ em 3 escalas modulados pelo filtro Gaussiano B.

4.5 RESULTADO COM FILTROS CONCATENADOS

Após o estudo realizado dos filtros isolados, pode-se observar que cada um deles consegue capturar de maneira distinta a informação de movimento contida nos vídeos. Com isso, a principal contribuição deste trabalho, é realizar uma combinação entre os descritores gerados, com objetivo de agrupar em apenas um descritor a capacidade de capturar os diversos movimentos ocorridos nos vídeos. Como dito na Seção 3.3.0.1, os melhores resultados encontrados foram obtidos usando a concatenação entre os descritores de cada vídeo. A Tabela 4.7 mostra a comparação entre duas possíveis combinações realizadas nos descritores, sendo elas: soma e concatenação. Vale ressaltar que essa soma ocorre entre os descritores obtidos dos filtros separadamente.

Filtros	**Somados**	**Concatenados**
$db1, db2$	90,9%	92,1%
$db1, db3$	89,3%	91,5%
$db1, db6$	91,8%	92,2%
$db2, db3$	86,7%	87,5%
$db1, db3, db7$	90,3 %	93,2%
$db1, db3, db8, db10$	89,7%	92,0%

Tabela 4.7: Taxa de reconhecimento para os tensores somados e concatenados.

A combinação dos descritores através da concatenação, mostra-se superior em relação à soma deles. É importante destacar que após a soma dos descritores, é realizada uma normalização no descritor final. A Figura 4.13 mostra um gráfico comparativo entre as combinações realizadas. Pode-se notar que a curva gerada pela soma dos descritores se mantém sempre abaixo da curva da concatenação entre eles. Outras combinações foram testadas, como por exemplo, a combinação no nível do histogramas de gradiente, porém essa e as demais não apresentaram um bom resultado. Assim, é proposto a concatenação como método de combinação de tensores.

Figura 4.13: Gráfico comparativo entre os filtros somados e concatenados.

A Tabela 4.8 mostra que o descritor gerado pela concatenação dos filtros $db1, db3, db7$ consegue realizar uma diferenciação entre os movimentos *Jog* e *Run* de forma mais satisfatória que o descritor do filtro $db1$ (Tab. 4.5).

	Box	HClap	HWav	Jog	Run	Walk
Box	95.8	2.8	0.00	0.00	0.00	1.4
HClap	2.1	95.8	2.1	0.00	0.00	0.00
HWav	6.2	0.00	93.8	0.00	0.00	0.00
Jog	0.7	0.00	0.00	90.3	6.2	2.8
Run	0.00	0.00	0.00	11.8	86.8	1.4
Walk	0.00	0.00	0.00	3.5	0.0	96.5

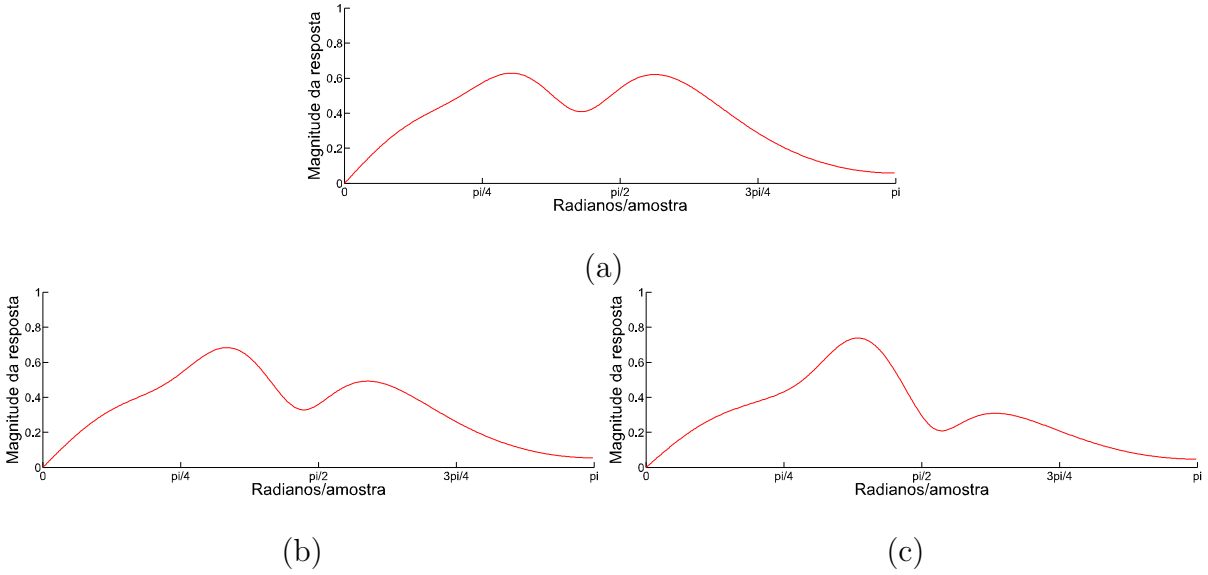
Tabela 4.8: Matriz de confusão para o filtro $db1, db3, db7$.

4.6 RESULTADO COM FILTROS CORRELACIONADOS

Com base nos estudos realizados de cada um dos filtros, pode-se afirmar que a ideia de combinar filtros distintos nos permite extrair diferentes tipos de movimento em uma

sequência de quadros. Cada filtro é capaz de capturar melhor alguns movimentos do que outros. Com isso, projeta-se um novo filtro com a finalidade de correlacionar os múltiplos espectros gerados por cada um deles.

Os filtros projetados neste trabalho foram baseados nos resultados obtidos na Tabela 4.7. A Figura 4.14 mostra a resposta de impulso para esses filtros.



(a)

(b)

(c)

Figura 4.14: Função de transferência dos filtros correlacionados modulado por uma gaussiana B nos eixos $\boldsymbol{x}$ e $\boldsymbol{y}$. (a) Correlação dos filtros $db1$, $db3$ e $db7$. (b) Correlação dos filtros $db1$, $db3$ e $db8$. (c) Correlação dos filtros $db1$, $db3$ e $db10$.

A proposta de projetar novos filtros, tem como objetivo encontrar a faixa do espectro que contém a maior quantidade de informação de movimento. Nota-se que os filtros de correlação obtêm taxas de reconhecimento próximas da média dos seus filtros constituintes. A Tabela 4.9 mostra os resultados obtidos por cada um desses filtros.

Filtro	**Taxa de reconhecimento**
$db_{1,3,7}$	85,5%
$db_{1,3,8}$	87,0%
$db_{1,3,10}$	86,3%

Tabela 4.9: Taxa de reconhecimento para os filtros correlacionados.

Como demonstrado na Seção 4.4.1, a concatenação dos filtros é eficaz para o aumento da taxa de reconhecimento. Portanto, realiza-se uma concatenação dos filtros projetados com os demais. Os resultados são mostrados na Tabela 4.10.

Filtro	Taxa de reconhecimento
$db1, db3, db7, db_{1,3,7}$	90,5%
$db1, db3, db8, db_{1,3,8}$	89,0%
$db1, db3, db10, db_{1,3,10}$	92,4%

Tabela 4.10: Taxa de reconhecimento para a concatenação dos filtros projetados.

O objetivo é usar o filtro projetado para correlacionar a resposta dos filtros individuais que o compõe.

A concatenação dos filtros individuais juntamente com o filtro que os correlaciona, aumenta a taxa de reconhecimento. Neste caso, a aplicação da normalização de energia, através de um fator γ, é feita somente no resultado do filtro de correlação. A Tabela 4.11 mostra os valores obtidos após essa normalização.

Filtro	Taxa de reconhecimento
$db1, db3, db7, db_{1,3,7}$ com $\gamma = 0,5$	93,3%
$db1, db3, db8, db_{1,3,8}$ com $\gamma = 0,5$	92,2%
$db1, db3, db10, db_{1,3,10}$ com $\gamma = 0,5$	92,6%

Tabela 4.11: Taxa de reconhecimento para a concatenação dos filtros projetados com normalização de energia.

Com a utilização da normalização de energia, percebe-se um aumento na taxa de reconhecimento dos filtros analisados. O filtro $db1, db3, db7, db_{1,3,7}$ com $\gamma = 0,5$ obteve o melhor resultado para o reconhecimento de ações humanas em vídeos. A Tabela 4.12 nos mostra a capacidade desse filtro para capturar cada um dos movimentos da base KTH.

	Box	HClap	HWav	Jog	Run	Walk
Box	95.8	2.8	0.00	0.00	0.00	1.4
HClap	2.1	96.5	1.4	0.00	0.00	0.00
HWav	6.2	0.00	93.8	0.00	0.00	0.00
Jog	0.7	0.00	0.00	90.3	6.2	2.8
Run	0.00	0.00	0.00	12.5	86.8	0.7
Walk	0.00	0.00	0.00	3.5	0.0	96.5

Tabela 4.12: Matriz de confusão para o filtro $db1, db3, db7, db_{1,3,7}$ com $\gamma = 0,5$.

4.7 COMPARAÇÃO COM OUTROS MÉTODOS PARA BASE KTH

Nesta seção, compara-se o melhor resultado encontrado com outros descritores na literatura. A Tabela 4.13 mostra o desempenho do método proposto, usando o filtro derivativo $db1, db3, db7, db_{1,3,7}$ com $\gamma = 0,5$.

Métodos globais	**Taxa de reconhecimento**
HOG pirâmidal (ZELNIK-MANOR; IRANI, 2001)	72.00%
Banco de filtros Gabor (SOLMAZ et al., 2012)	92.00%
HOG3D + Tensor (PEREZ et al., 2012)	92.01%
Método Proposto (4 filtros)	**93.30%**
Métodos locais	**Taxa de reconhecimento**
Harris3D + HOG/HOF (LAPTEV et al., 2008)	91.80%
Pontos de interesse + Wavelets (SHAO; GAO, 2010)	93.89%
HOG+HOF+MBH+Trajetória (WANG et al., 2011)	94.20%
DT-CWT+SIFT (MINHAS et al., 2010)	94.83%

Tabela 4.13: Comparação com outros métodos para base KTH.

Comparando o descritor global proposto nesta dissertação com os demais, é possível dizer que a metodologia de concatenar descritores, gerados por filtros distintos, é eficaz para o reconhecimento de ações humanas em vídeos. Pode-se observar que este método apresenta um resultado competitivo se comparado aos métodos locais (LAPTEV et al., 2008; SHAO; GAO, 2010), com a vantagem de ser muito mais simples, necessitando de baixo poder computacional. Outros métodos globais, como por exemplo, o descritor apresentado em Solmaz et al. (2012), além de utilizar um banco com 68 filtros de Gabor, utiliza uma técnica de redução de dimensionalidade conhecida como Análise de Componentes Principais. O melhor resultado alcançado pelo descritor proposto neste trabalho, utiliza apenas 4 filtros e atinge um resultado superior aos demais (Tab. 4.13).

5 CONCLUSÃO

Neste trabalho, foi apresentado uma nova abordagem para a descrição de movimento em vídeos, através da concatenção de vários filtros. Esses filtros agem como estimadores derivativos de primeira ordem. Essa abordagem se mostra eficaz, pois consegue atingir $93,3\%$ de taxa de reconhecimento na base KTH, superando outros métodos globais e sendo competitiva se comparada aos métodos locais e de aprendizagem como mostra a Tabela 4.13. Além disso, o descritor proposto apresenta uma abordagem muito mais simples, usando apenas informações extraídas pelos filtros derivativos, sem o uso da estratégia conhecida como dicionário de característica (*bag of features*)(LAPTEV et al., 2008; SHAO; GAO, 2010; WANG et al., 2011). Para criação do descritor, realizou-se um estudo comparativo entre os melhores resultados obtidos por cada um dos filtros apresentados neste trabalho. Foi observado que o filtro $db1$ sempre apresentou altas taxas de reconhecimento, mesmo quando combinado com outros filtros (Tab. 5.1).

Filtro	**Taxa de reconhecimento**
$db1$	90,9%
$db1, db3$	91,5%
$db1, db7$	92,6%
$db1, db3, db7$	93,2%
$db1, db3, db7, db_{1,3,7}$ com $\gamma = 0,5$	93,3%

Tabela 5.1: Taxa de reconhecimento usando o filtro $db1$.

Com base nos resultados encontrados, observou-se que a concatenação entre os descritores gerados por cada um dos filtros, é uma abordagem válida para classificar a base de dados KTH. O uso da normalização de energia dos gradientes proporcionou um aumento na taxa de classificação, sendo visível principalmente em ações com movimentos mais abrutos, como o *running*, *hand clapping* e *hand waving*.

Alguns autores utilizam outras técnicas de classificação, por exemplo, o protocolo *leave-one-out* (MINHAS et al., 2010). Apesar de apresentar uma investigação completa sobre a variação do modelo em relação aos dados utilizados, este protocolo possui um alto custo computacional, sendo indicado para situações onde poucos dados estão disponíveis. Usando este protocolo, o método proposto alcança $95,5\%$ de taxa de reconhecimento usando o filtro $db1, db3, db10$. Os resultados aqui apresentados, indicam que o estudo

dos filtros derivativos que melhor conseguem extrair informações sobre um determinado movimento é promissor para o problema de reconhecimento de ações humanas em vídeos.

Alguns descritores foram gerados para classificar os vídeos da base de dados Hollywood2 (MARSZAŁEK et al., 2009). É possível observar que o filtro $db1$ isoladamente consegue a melhor taxa de reconhecimento nesta base, assim como na KTH, porém, a concatenação entre alguns filtros não apresentou uma melhora nos resultados. Portanto, podemos concluir que, para cada base de vídeo utilizada, é necessário investigar qual a melhor combinação de filtros que deve ser utilizada para obter uma boa taxa de reconhecimento (Tab. 5.2).

Filtro	**Taxa de reconhecimento**
$db1$	41,9%
$db2$	34,4%
$db3$	30,5%
$db1, db3$	41,9%
$db1, db2, db3$	41,2%

Tabela 5.2: Taxa de reconhecimento para a base Hollywood2.

Para trabalhos futuros, é necessário aprofundar o estudo dos filtros derivativos, analisando sua capacidade de extrair cada um dos movimentos realizados em um vídeo. Outro ponto a ser estudado, é em relação a qual filtro suavizador deve ser utilizado, uma vez que ele modifica substancialmente todos os filtros derivativos que são aplicados em cada quadro do vídeo.

Uma possível aplicação do uso de múltiplos filtros para extração de movimento, está relacionada ao reconhecimento de uma pessoa através do movimento característico daquele indivíduo. Nos últimos anos, a biometria se mostra como uma tecnologia segura e robusta para este fim. Os sistemas biométricos atuais são geralmente baseados em apenas uma característica do indivíduo, o que dificulta o reconhecimento. Para minimizar esses problemas e melhorar as taxas de identificação, têm sido propostas técnicas de multibiometria, ou seja, uma combinação de evidências biométricas (SANDERSON; PALIWAL, 2003). Uma das características biométricas que podem ser analisadas para aumentar a taxa de reconhecimento de indivíduos é através do estudo dos movimentos característicos dessa pessoa.

REFERÊNCIAS

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: SCHMID, C.; SOATTO, S.; TOMASI, C. (Ed.). **International Conference on Computer Vision & Pattern Recognition**, 2005. v. 2, p. 886–893. Disponível em: <http://lear.inrialpes.fr/pubs/2005/DT05>.

FOURNIER, J.; CORD, M.; PHILIPP-FOLIGUET, S. RETIN: A Content-Based Image Indexing and Retrieval System. **Pattern Analysis & Applications**, v. 4, n. 2, p. 153–173, June 2001. ISSN 1433-7541. Disponível em: <http://dx.doi.org/10.1007/PL00014576>.

GORELICK, L.; BLANK, M.; SHECHTMAN, E.; IRANI, M.; BASRI, R. Actions as space-time shapes. In: **In ICCV**, 2005. p. 1395–1402.

HAYKIN, S. **Redes Neurais - 2ed.**, 2001. ISBN 9788573077186. Disponível em: <http://books.google.com.br/books?id=lBp0X5qfyjUC>.

JOHANSSON, B.; FARNEBCK, G.; ACK, G. F. A theoretical comparison of different orientation tensors. In: **Symposium on Image Analysis**, 2002. p. 69–73.

KHADEM, B. S.; RAJAN, D. Appearance-based action recognition in the tensor framework. In: **Proceedings of the 8th IEEE international conference on Computational intelligence in robotics and automation**, 2009. (CIRA'09), p. 398–403. ISBN 978-1-4244-4808-1. Disponível em: <http://dl.acm.org/citation.cfm?id=1811259.1811340>.

KIHL, O.; TREMBLAIS, B.; AUGEREAU, B.; KHOUDEIR, M. Human activities discrimination with motion approximation in polynomial bases. In: **IEEE International Conference on Image Processing**, 2010. p. 2469–2472. Disponível em: <http://hal.archives-ouvertes.fr/hal-00594762/en/>.

KIM, T.; WONG, S.; CIPOLLA, R. R.: Tensor canonical correlation analysis for action classification. In: **In: CVPR 2007**, 2007.

KLÄSER, A.; MARSZAŁEK, M.; SCHMID, C. A spatio-temporal descriptor based on 3d-gradients. In: **British Machine Vision Conference**, 2008. p. 995–1004. Disponível em: <http://lear.inrialpes.fr/pubs/2008/KMS08>.

KRAUSZ, B.; BAUCKHAGE, C. Action recognition in videos using nonnegative tensor factorization. In: **Proceedings of the 2010 20th International Conference on Pattern Recognition**, 2010. (ICPR '10), p. 1763–1766. ISBN 978-0-7695-4109-9. Disponível em: <http://dx.doi.org/10.1109/ICPR.2010.435>.

LAPTEV, I.; CAPUTO, B.; SCHULDT, C.; LINDEBERG, T. Local velocity-adapted motion events for spatio-temporal recognition. **Comput. Vis. Image Underst.**, Elsevier Science Inc., New York, NY, USA, v. 108, p. 207–229, December 2007. ISSN 1077-3142.

LAPTEV, I.; MARSZAŁEK, M.; SCHMID, C.; ROZENFELD, B. Learning realistic human actions from movies. In: **Computer Vision & Pattern Recognition**, 2008. Disponível em: <http://lear.inrialpes.fr/pubs/2008/LMSR08>.

LOWE, D. G. Object recognition from local scale-invariant features. In: **Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2**, 1999. (ICCV '99), p. 1150–. ISBN 0-7695-0164-8. Disponível em: <http://dl.acm.org/citation.cfm?id=850924.851523>.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **Int. J. Comput. Vision**, Kluwer Academic Publishers, Hingham, MA, USA, v. 60, n. 2, p. 91–110, nov 2004. ISSN 0920-5691. Disponível em: <http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94>.

MALLAT, S. **A wavelet tour of signal processing (2. ed.)**, 1999. I-XXIV, 1-637 p. ISBN 978-0-12-466606-1.

MARR, D.; POGGIO, T.; ULLMAN, S. **Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information**, 2010. ISBN 9780262514620. Disponível em: <http://books.google.com.br/books?id=EehUQwAACAAJ>.

MARSZAŁEK, M.; LAPTEV, I.; SCHMID, C. Actions in context. In: **Conference on Computer Vision & Pattern Recognition**, 2009. Disponível em: <http://lear.inrialpes.fr/pubs/2009/MLS09>.

MILIC, L. **Multirate filtering for digital signal processing : MATLAB applications / Ljiljana Milic.**, 2009.

MINHAS, R.; BARADARANI, A.; SEIFZADEH, S.; WU, Q. J. Human action recognition using extreme learning machine based on visual vocabularies. **Neurocomputing**, v. 73, 2010. ISSN 0925-2312. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231210001517>.

MOTA, V. F. **Tensor baseado em fluxo óptico para descrição global de movimento em vídeos**. Dissertação (Mestrado) — Universidade Federal de Juiz de Fora, 2011.

PEREZ, E. A. **Descritor de movimento baseado em tensor e histograma de gradientes**. Dissertação (Mestrado) — Universidade Federal de Juiz de Fora, 2012.

PEREZ, E. A.; MOTA, V. F.; MACIEL, L. M.; SAD, D.; VIEIRA, M. B. Combining gradient histograms using orientation tensors for human action recognition. In: **International Conference on Pattern Recognition**, 2012.

SANDERSON, C.; PALIWAL, K. K. **Noise Compensation in a Person Verification System Using Face and Multiple Speech Features**. 2003.

SCHULDT, C.; LAPTEV, I.; CAPUTO, B. Recognizing human actions: A local svm approach. In: **In Proc. ICPR**, 2004. p. 32–36.

SHAO, L.; GAO, R. A wavelet based local descriptor for human action recognition. In: **Proc. BMVC**, 2010. p. 72.1–10. ISBN 1-901725-40-5. Doi:10.5244/C.24.72.

SMOLA, A. J.; BARTLETT, P. J. (Ed.). **Advances in Large Margin Classifiers**, 2000. ISBN 0262194481.

SOLMAZ, B.; ASSARI, S. M.; SHAH, M. Classifying web videos using a global video descriptor. **Machine Vision and Applications**, Springer Berlin / Heidelberg, p. 1–13,

sep 2012. ISSN 0932-8092. Disponível em: <http://dx.doi.org/10.1007/s00138-012-0449-x>.

SUNG, A.; MUKKAMALA, S. Identifying important features for intrusion detection using support vector machines and neural networks. In: **Applications and the Internet, 2003. Proceedings. 2003 Symposium on**, 2003. p. 209 – 216.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**, 1995.

WANG, H.; KLÄSER, A.; SCHMID, C.; CHENG-LIN, L. Action Recognition by Dense Trajectories. In: **IEEE Conference on Computer Vision & Pattern Recognition**, 2011. p. 3169–3176. Disponível em: <http://hal.inria.fr/inria-00583818>.

WESTIN, C.-F. **A Tensor Framework for Multidimensional Signal Processing**. Tese (Doutorado) — Linköping University, Sweden, 1994. N. 348.

ZELNIK-MANOR, L.; IRANI, M. Event-based analysis of video. In: **In Proc. CVPR**, 2001. p. 123–130.

Apêndice A - COEFICIENTES DOS FILTROS

Daubechies 1

Passa alta = {-0.70, 0.70}

Passa baixa = {0.70, 0.70}

Daubechies 2

Passa alta = {-0.48, 0.83, -0.22, -0.12}

Passa baixa = {-0.12, 0.22, 0.83, 0.48}

Daubechies 3

Passa alta = {-0.33, 0.80, -0.45, -0.13, 0.08, 0.03}

Passa baixa = {0.03, -0.08, -0.13, 0.45, 0.80, 0.33}

Daubechies 4

Passa alta = {-0.23, 0.71, -0.63, -0.02, 0.18, 0.03, -0.03, -0.01}

Passa baixa = {-0.01, 0.03, 0.03, -0.18, -0.02, 0.63, 0.71, 0.23}

Daubechies 5

Passa alta = {-0.16, 0.60, -0.72, 0.13, 0.24, -0.03, -0.07, 0.00, 0.01, 0.00}

Passa baixa = {0.00, -0.01, 0.00, 0.07, -0.03, -0.24, 0.13, 0.72, 0.60, 0.16}

Daubechies 6

Passa alta = {0.00, 0.00, 0.00, -0.03, 0.02, 0.09, -0.12, -0.22, 0.31, 0.75, 0.49, 0.11}

Passa baixa = {-0.11, -0.75, 0.31, 0.22, -0.12, -0.09, 0.02, 0.03, 0.00, 0.00, 0.00}

Daubechies 7

Passa alta = {-0.07, 0.39, -0.72, 0.46, 0.14, -0.22, -0.07, 0.08, 0.03, -0.01, -0.01, 0.00, 0.00, 0.00}

Passa baixa = {0.00, 0.00, 0.00, 0.01, -0.01, -0.03, 0.08, 0.07, -0.22, -0.14, 0.46, 0.72, 0.39, 0.07}

Daubechies 8

Passa alta = {-0.05, 0.31, -0.67, 0.58, 0.01, -0.28, 0.00, 0.12, 0.01, -0.04, -0.01, 0.00, 0.00, 0.00, 0.00, 0.00}

Passa baixa = {0.00, 0.00, 0.00, 0.00, 0.00, 0.01, -0.04, -0.01, 0.12, 0.00, -0.28, -0.01, 0.58, 0.67, 0.31, 0.05}

Daubechies 9

Passa alta = {-0.03, 0.24, -0.60, 0.65, -0.13, -0.29, 0.09, 0.14, -0.03, -0.06, 0.00, 0.02, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00}

Passa baixa = {0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.02, 0.00, -0.06, 0.03, 0.14, -0.09, -0.29, 0.13, 0.65, 0.60, 0.24, 0.03}

Daubechies 10

Passa alta = {-0.02, 0.18, -0.52, 0.68, -0.28, -0.24, 0.19, 0.12, -0.09, -0.07, 0.02, 0.03, 0.00, -0.01, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00}

Passa baixa = {0.00, 0.00, 0.00, 0.00, 0.00, 0.00, -0.01, 0.00, 0.03, -0.02, -0.07, 0.09, 0.12, -0.19, -0.24, 0.28, 0.68, 0.52, 0.18, 0.02}

Sobel

Passa alta = {-0.50, 0.00, 0.50}

Passa baixa = {0.50, 1.00, 0.5}

Coiflets 1

Passa alta = {0.07, 0.33, -0.85, 0.38, 0.07, -0.01}

Passa baixa = {-0.01, -0.07, 0.38, 0.85, 0.33, -0.07}

Coiflets 2

Passa alta = {-0.01, -0.04, 0.06, 0.38, -0.81, 0.41, 0.07, -0.05, -0.02, 0.00, 0.00, 0.00}

Passa baixa = {0.00, 0.00, 0.00, 0.02, -0.05, -0.07, 0.41, 0.81, 0.38, -0.06, -0.04, 0.01}

Symlets 2

Passa alta = {-0.48, 0.83, -0.22, -0.12}

Passa baixa = {-0.12, 0.22, 0.83, 0.48}

Biorthogonal 1.3

Passa alta = {0.0, 0.0, -0.70, 0.70, 0.0, 0.0}

Passa baixa = {-0.08, 0.08, 0.70, 0.70, 0.08, -0.08}

$db_{1,3,7}$

Passa alta = {-0.21, 0.36, -0.22, 0.06, 0.04, -0.03, -0.01, 0.01}

Passa baixa = {0.21, 0.63, 0.22, -0.06, -0.04, 0.03, 0.01, -0.01}

$db_{1,3,8}$

Passa alta = {-0.25,0.43, -0.26, 0.10, 0.02, -0.05, 0.00, 0.03, 0.00, -0.01}

Passa baixa = {0.25, 0.56, 0.26, -0.10, -0.02, 0.05, 0.00, -0.03, 0.00, 0.01}

$db_{1,3,10}$

Passa alta = { -0.75, 0.20, -0.69, 0.39, -0.13, -0.15, 0.13, 0.09, -0.06, -0.05, 0.02, 0.02}

$$Passa\ baixa = \{0.00,\ 0.00,\ 0.00,\ 0.00,\ 0.00,\ 0.00,\ 0.00,\ 0.25,\ 0.50,\ 0.25\}$$