Virgínia Fernandes Mota

Tensor baseado em fluxo óptico para descrição global de movimento em vídeos

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre em Modelagem Computacional.

Orientador: Prof. D.Sc. Marcelo Bernardes Vieira

Juiz de Fora

2011

Mota, Virgínia Fernandes.

Tensor baseado em fluxo óptico para descrição global de movimento em vídeos/ Virgínia Fernandes Mota. – 2011.
179 f. : il.

Dissertação (Mestrado em Modelagem Computacional)- Universidade Federal de Juiz de Fora, Juiz de Fora, 2011.

1. Ciência da computação. 2. Descritor de movimento. 3. Tensor de orientação. 4. SVM. 5. Fluxo Óptico. 6. Modelagem do movimento. I. Título.

CDU 681.3

Virgínia Fernandes Mota

Tensor baseado em fluxo óptico para descrição global de movimento em vídeos

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Mestre em Modelagem Computacional.

Aprovada em 28 de Fevereiro de 2011.

BANCA EXAMINADORA

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

Prof. D.Sc. Anselmo Antunes Montenegro
Universidade Federal Fluminense

Prof. D.Sc. Carlos Cristiano Hasenclever Borges
Universidade Federal de Juiz de Fora

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

Dedico este trabalho ao meu namorado Tiago Machado e à minha família pelo apoio e amor incondicionais.

AGRADECIMENTOS

Agradeço ao meu orientador Marcelo Bernardes Vieira por sua dedicação, companheirismo e paciência. Agradeço também aos meus orientadores da ENSEA Philippe-Henri Gosselin, Sylvie Philipp-Foliguet e, principalmente, Frédéric Precioso, sem ele este trabalho não seria possível.

Aos membros da banca por terem aceitado o convite e por suas contribuições.

Agradeço aos professores do MMC que me ajudaram durante essa caminhada e também a todos os companheiros de estudo, principalmente, minha grande amiga Anna Paula Guida Ferreira por tornar o tempo que estudei na França muito mais agradável. Agradeço também a todos os meus companheiros de estudo na França, principalmente minha grande amiga Leila Meziou por toda sua ajuda e por suas aulas de francês. Aos membros do GCG, agradeço pela diversão que é trabalhar com todos.

Agradeço a todos os meus amigos pelo apoio, principalmente, meu grande amigo Rafael Ribeiro de Carvalho, que mesmo com problemas sempre esteve disposto a me ajudar.

Ao meu namorado, Tiago Machado, pelo apoio e incentivo de sempre. À minha mãe, Natália Maria da Silva Fernandes, e todos os meus familiares.

Por fim, agradeço a todos aqueles que de alguma forma fizeram parte dessa trajetória, mas não foram citados aqui, as palavras faltam para agradecê-los.

"O espelho e os sonhos são coisas semelhantes, é como a imagem do homem diante de si próprio."

José Saramago

RESUMO

Movimento é uma das características fundamentais que refletem a informação semântica em vídeos. Uma das técnicas de estimativa do movimento é o cálculo do fluxo óptico. Este é uma representação 2D (bidimensional) das velocidades aparentes de uma sequência de quadros (frames) adjacentes, ou seja, a projeção 2D do movimento 3D (tridimensional) projetado na câmera.

Neste trabalho é proposto um descritor global de movimento baseado no tensor de orientação. O mesmo é formado à partir dos coeficientes dos polinômios de Legendre calculados para cada quadro de um vídeo. Os coeficientes são encontrados através da projeção do fluxo óptico nos polinômios de Legendre, obtendo-se uma representação polinomial do movimento.

O descritor tensorial criado é avaliado classificando-se a base de vídeos KTH com um classificador SVM (máquina de vetor de suporte). É possível concluir que a precisão da abordagem deste trabalho supera às encontradas pelos descritores globais encontrados na literatura.

Palavras-chave: Descritor de movimento. Tensor de orientação. SVM. Fluxo Óptico. Modelagem do movimento.

ABSTRACT

Motion is one of the main characteristics that describe the semantic information of videos. One of the techniques of motion estimation is the extraction of optical flow. The optical flow is a bidimensional representation of velocities in a sequence of adjacent frames, in other words, is the 2D projection of the 3D motion projected on the camera.

In this work it is proposed a global video descriptor based on orientation tensor. This descriptor is composed by coefficients of Legendre polynomials calculated for each video frame. The coefficients are found though the projection of the optical flow on Legendre polynomials, obtaining a polynomial representation of the motion.

The tensorial descriptor created is evaluated by a classification of the KTH video database with a SVM (support vector machine) classifier. Results show that the precision of our approach is greater than those obtained by global descriptors in the literature.

Keywords: Motion descriptor. Orientation tensor. SVM. Optical Flow. Motion modeling.

SUMÁRIO

1 INTRODUÇÃO

A pesquisa em Visão Computacional define algoritmos permitindo a percepção e a compreensão do mundo físico à partir de informações visuais como imagens e sequências de imagens (vídeos). Um sistema de visão completo pode ser dividido em três etapas: extração de atributos de baixo-nível, tais como: cor, textura, forma, orientação e movimento; análise de atributos fornecendo informações de mais alto-nível, tais como reconhecimento, segmentação e classificação; e interpretação da cena. Este trabalho trata das duas primeiras etapas.

Movimento é uma das características fundamentais que reflete a informação semântica em vídeos. A habilidade de detectar um objeto e/ou uma pessoa e rastreá-lo é de grande interesse em várias aplicações de segurança, como por exemplo rastreamento de mísseis e detecção de movimento em sistemas de vigilância.

Uma das técnicas de estimativa do movimento é o cálculo do fluxo óptico. O fluxo óptico é uma estimativa da representação bidimensional das velocidades aparentes de uma sequência de quadros adjacentes, ou seja, a representação do movimento projetado na câmera. O cálculo do movimento entre imagens de uma sequência corresponde à estimação dos parâmetros de uma transformação de todos os pontos da imagem: translação, rotação, lineares e afins, e deformações diversas. Em uma cena real, as transformações de cada ponto são arbitrariamente complexas. O processo de inferência de movimento geralmente lida com correlações que possuem natureza multivariada. Com isto, esse contexto se torna propício ao uso de tensores.

Uma das abordagens atuais para a extração do fluxo óptico é utilizar o tensor de estrutura. O mesmo descreve uma informação média local das orientações dos vetores de velocidade e preserva a estrutura do movimento [1]. Estes métodos baseados em tensores geralmente estimam o movimento a partir de métodos diferenciais e de filtragem, já que o movimento pode ser caracterizado como a variação instantânea da luminância dos *pixels* entre imagens adjacentes. O uso de tensores possibilita capturar a covariância dos métodos básicos descrevendo a estrutura espaço-temporal de uma dada vizinhança. Além disso, fornece informação sobre a velocidade local e se o fluxo verdadeiro ou o fluxo normal está presente [2],[3].

A partir da extração do movimento do vídeo, é possível passar para a segunda etapa do sistema de visão, a etapa de análise dos atributos. Neste trabalho o tipo de análise de atributos tratado é o reconhecimento de ações.

1.1 Definição do problema e objetivos

A necessidade de reconhecimento de ações surgiu dos sistemas de Extração da Informação de Imagens e Vídeos Baseada em Conteúdo (CBIVR - do Inglês *Content-Based Image and Video Retrieval*). Estes tipos de sistema consideram a extração de um conjunto de características associadas a cada tomada da sequência e a sua posterior representação em descritores. Estes descritores são usados em ferramentas de busca, comparação e classificação das respectivas tomadas. Sistemas como estes tem sido usados em diversas outras aplicações tais como: identificação de digitais [4], aplicações médicas [5], reconstrução de fachadas [6], reconhecimento de gestos [7], sistemas de segurança [8], entre outras.

O reconhecimento de ações humanas em vídeos tem diversas aplicações na área de segurança, entretenimento, interface homem-máquina, entre outros domínios. Dado um número pré-definido de ações, o problema pode ser definido como classificar uma nova ação dentre essas ações. Geralmente, este conjunto de ações possui relevância para um certo domínio. Por exemplo, na linguagem de sinais, um conjunto de ações representa um conjunto possível de palavras e letras que podem ser produzidas.

Um método interessante para a análise do movimento é a modelagem por combinações lineares de polinômios ortogonais. Este método permite obter uma expressão polinomial do movimento [3], [9].

O objetivo deste trabalho é apresentar um novo descritor global baseado no tensor de orientação para o problema de reconhecimento de ações em vídeos. Este tensor é composto à partir dos coeficientes dos polinômios de Legendre calculados para cada tomada de um vídeo.

A principal contribuição deste trabalho está na aproximação do fluxo óptico por base de polinômios de Legendre, codificá-la em um tensor de orientação e acumular estes tensores para a criação de um descritor global.

Para avaliar o descritor é utilizada a base de vídeos KTH [10] na qual existem seis tipos de ações humanas (*walking*, *jogging*, *running*, *boxing*, *hand waving* e *hand clapping*)

executadas diversas vezes por 25 pessoas diferentes e em quatro cenários: ambiente externo (s1), ambiente externo com variação de escala (s2), ambiente externo com variação de velocidade (s3) e ambiente interno (s4) (Figura 1.1). Todas as 2391 sequências são realizadas com fundo homogêneo e uma câmera estática de 25 quadros por segundo. As sequências tem uma resolução de 160x120 *pixels* e duram em torno de quatro segundos.

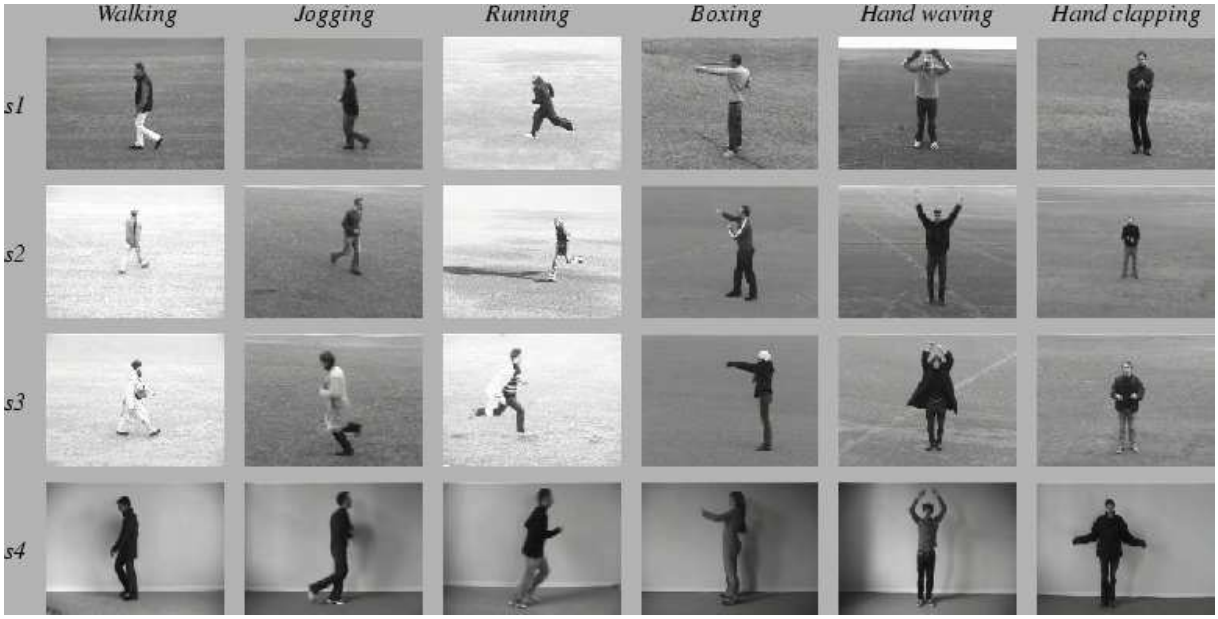


Figura 1.1: Tipos de ações da base de vídeos KTH [10]

Esta base é classificada utilizando um classificador SVM (máquina de vetor de suporte) e a precisão encontrada é utilizada para qualificar o descritor proposto.

1.2 Organização do trabalho

Este trabalho está dividido em cinco partes. No Capítulo 2, são apresentados alguns trabalhos relacionados à esta dissertação e são divididos em duas categorias principais: cálculo do fluxo óptico e análise do movimento.

No Capítulo 3, são apresentados os fundamentos mais importantes para a compreensão deste trabalho. São eles: o fluxo óptico, a modelagem de campos de deslocamento por base de polinômios, o tensor de orientação e os descritores globais utilizados para comparação com o descritor proposto (o descritor de Zelnik *et al* [11] e o de Laptev *et al* [12]).

No Capítulo 4, é apresentada a proposta deste trabalho: a criação de um descritor global de movimento baseado no tensor de orientação e algumas variações propostas. Os

resultados obtidos aplicando o descritor global na classificação de vídeos são apresentados e discutidos no Capítulo 5.

No último capítulo, é apresentada uma conclusão geral do trabalho, destacando os pontos fortes e fracos do descritor e são propostos alguns trabalhos futuros.

2 Trabalhos Relacionados

Neste capítulo, são mostrados diferentes métodos que permitem a extração do movimento contido em uma sequência de imagens e sua análise.

A Seção 2.1 apresenta alguns métodos para a extração do movimento. Na Seção 2.2 são apresentados diversos métodos para a análise do movimento bem como alguns métodos baseados em tensores.

2.1 Cálculo do Fluxo Óptico

O artigo de Lauze *et al* [2] trata da estimativa do fluxo óptico à partir da análise do campo de tensores de estrutura de uma sequência de imagens vista como um volume espaço-temporal. A estrutura do volume reagrupa a informação espacial das imagens por um tempo dado, bem como sua evolução temporal. Um método para analisar esta estrutura é calcular o tensor de estrutura. Os autores descrevem três abordagens para encontrar este tipo de tensor: o método de Lucas e Kanade [13], a análise de autovalores e autovetores e a utilização do fluxo normal e medidas de coerência.

O método de Lucas e Kanade é uma simples estimativa de médias ponderadas para o sistema de restrições do fluxo óptico na vizinhança de (x, t). Este método possui grande utilização por ser de fácil implementação e ser numericamente estável. A análise dos autovalores e autovetores é também uma estimativa de médias ponderadas para o sistema de restrições do fluxo óptico, porém menos estável numericamente que a anterior. Já a utilização do fluxo normal pressupõe a extração do fluxo normal em uma vizinhança de (x, t) e estima o fluxo óptico resolvendo o sistema de restrições projetando o fluxo óptico no fluxo normal. É um método melhor que a análise de autovalores e autovetores, mas não melhor que o método de Lucas e Kanade.

Horn e Schunck [14] calculam o fluxo óptico à partir das derivadas espaço-temporais das intensidades na imagem. Para evitar variações no brilho devido a efeitos de sombras, é assumido que a superfície a ser trabalhada é plana e também que a iluminação incidente sobre a superfície é uniforme. A grande diferença deste método é que ele acrescenta ao cálculo do fluxo óptico uma restrição de coerência espacial do campo de deslocamento.

O método de Augereau [1] calcula o fluxo óptico através de um método diferencial de estimação do movimento aparente. Este método é fundado na hipótese da conservação da luminância dos *pixels* de uma imagem e na utilização de equações de derivadas parciais de filtragem direcional. Estas utilizam um operador diferencial chamado tensor de estrutura, permitindo determinar localmente, a partir de dados espaço-temporais, a direção do movimento aparente. Assim, uma avaliação do campo denso e regularizado é obtida. A diferença entre este método e o de Lucas e Kanade está no uso das características espectrais do tensor de estrutura. Além disso, esse método encontra um fluxo óptico mais regular que o encontrado pelo método de Lucas e Kanade.

Os métodos de Lucas e Kanade, Horn e Schunck e Augereau são considerados métodos diferenciais de extração de movimento e tem por princípio estudar a variação temporal das intensidades luminosas na sequência de imagens [15]. Para tal, estes assumem que a quantidade total de intensidades luminosas não varia entre quadros (*frames*) adjacentes do vídeo. Dessa forma, mesmo sendo considerado uma boa aproximação do movimento aparente, o cálculo do fluxo óptico é sensível a ruído e às mudanças de iluminação.

2.2 Análise do Movimento

A abordagem de Hayko *et al* [16] considera a sequência de imagens como um volume espaço-temporal e detecta os *Maximally Stable Volumes* (MSVs) nos campos de fluxo óptico, que são volumes estáveis do fluxo óptico. O método para identificar as sequências de imagens é composto de duas etapas: deteção e descrição dos pontos de interesse 3D (*feature vectors*) e o modelo de *bag-of-words* que descreve as sequências de imagens em termos de assinatura do volume do fluxo óptico.

A detecção e descrição de pontos de interesse se compõe principalmente de três etapas: estimação do fluxo óptico (pelo método TV-L1 [17]), aplicação do detector *Maximally Stable Volume* (MSV) para identificar os volumes estáveis do fluxo óptico e a descrição dos pontos de interesse situados na superfícies do volume. As *visual words* são identificadas agrupando o conjunto de descritores para caracterizar suas propriedades semelhantes. A partir do número de ocorrências das *visual words* na sequência de imagens, uma assinatura distinta é construída.

Em [18], é apresentada uma comparação entre várias abordagens para o reconheci-

mento de ações humanas utilizando a técnica de *bag-of-visual-features*. São avaliadas a taxa de reconhecimento e a complexidade de dois descritores 2D e um descritor 3D, respectivamente, SIFT (*Scale-Invariant Features Transform*) [19], SURF (*Speed-Up Robust Features*) [20] e STIP (*Spatio-Temporal Interest Points*) [21]. SIFT é um detector de pontos de interesse e descritor que procura por pontos que apresentam invariância em relação à posição, escala e localização. O algoritmo SURF possui os mesmos objetivos do SIFT, porém é modificado para uma melhor performance. STIP é uma extensão do detector de Harris no espaço 3D composto por quadros e a dimensão do tempo. As características procuradas por este detector maximizam a variação dos gradientes em níveis de cinza. Os resultados mostraram que os descritores SIFT tiveram uma melhor performance em relação aos descritores SURF e os descritores 2D encontraram a mesma taxa de reconhecimento do descritor 3D, entretanto com uma maior complexidade. Uma comparação mais aprofundada entre SIFT e SURF pode ser encontrada em [22].

Em [3], o método apresentado propõe caracterizar todo tipo de movimento como uma combinação linear de polinômios de uma base ortonormal. A partir da sequência original, todos os campos de vetores do fluxo óptico são extraídos pelo método de Augereau [1]. Para cada um destes campos, são calculados os polinômios característicos através de projeções na base. Finalmente, o movimento é determinado estudando as variáveis dos coeficientes destes polinômios no tempo. Neste trabalho, são estudados movimentos simples da cabeça (verticais e horizontais).

Um estudo mais aprofundado da modelagem do movimento por base de polinômios é proposto em [9]. Diferente de [3], o método é aplicado para o escoamento de fluidos e permite modelar, de forma global, todo tipo de movimento através de combinações lineares de polinômios ortogonais. Este permite obter uma expressão polinomial do movimento estudado, assim, analisando os coeficientes destes polinômios é possível obter o comportamento deste movimento e extrair diversas informações.

Os artigos de Hayko *et al* [16] e os de Lopes *et al* ([18] e [22]) tratam do reconhecimento de ações humanas da base de vídeo Weizmann [23]. Já os trabalhos de Druon *et al* [3] e Druon [9] tratam da análise do movimento da cabeça [3] e do movimento de fluidos [9] a partir do comportamento dos coeficientes dos polinômios no tempo. Os próximos artigos tratam do reconhecimento de ações e trabalham sobre a base de vídeo KTH, introduzida por [10].

O método proposto por [10] utiliza o descritor *Spatio-Temporal Interest Points* (STIP). A partir da extração dos pontos de interesse, é utilizado um classificador SVM (máquinas de vetores de suporte - do inglês *support vectors machines*) para determinar a ação de cada vídeo. A precisão encontrada por este método, isto é, a porcentagem de vídeos classificados corretamente, foi de 71,7%.

Em [24], tem-se uma representação compacta para o reconhecimento de ações humanas em vídeos utilizando histogramas de linha e histogramas de fluxo óptico. O descritor contido nesse trabalho é baseado na distribuição das linhas formadas pelo contorno da figura humana combinado a uma representação compacta do fluxo óptico. A partir da extração destas características é utilizado um classificador SVM, tal como apresentado em [10]. A precisão apresentada neste artigo é de 94%.

Em [25], é proposto um método que combina um descritor 3D de gradiente com um descritor de fluxo óptico, o qual representa pontos de interesse espaço-temporais. Estes pontos são utilizados para representar as sequências de imagens com ajuda de *visual words*, de forma similar à [16]. Os classificadores SVM são também utilizados para a classificação da base e encontram uma precisão de 91,2%.

Em contraste com os outros estudos de reconhecimento de ações, Baysal *et al* [26] não utiliza a informação temporal para criar seu descritor. O descritor é criado utilizando posturas chaves dos humanos que representam cada tipo de ação. Para extrair as posturas chaves cada quadro é classificado pelo seu potencial em distinguir uma ação das outras. Posterior a isto são escolhidos os K melhores quadros que representam aquela ação e estes serão as posturas chaves desta ação. Para classificar uma dada ação, cada quadro da sequência é comparado com todas as posturas chaves de todas as ações e é rotulado com a ação cuja postura chave é mais similar. A rotulação final é feita escolhendo entre os rótulos atribuídos aquele que mais aparece. Esta abordagem alcança a precisão de 91,5%.

Em geral, os descritores para o reconhecimento de ações em vídeos que são compostos por mais de um tipo de característica alcançam melhores resultados. Estes descritores combinam características locais com características globais (como por exemplo o movimento). Existem poucas referências para descritores puramente globais.

Zelnik *et al* [11] apresenta um descritor global baseado em histogramas de gradientes. Esse descritor é aplicado à base de vídeo Weizmann [23] e é obtido extraindo-se características em múltiplas escalas temporais através da construção de uma pirâmide

temporal. O cálculo da mesma se dá por uma filtragem passa-baixa e amostragem da sequência na direção do tempo. Para cada uma das sequências obtidas, é calculado o gradiente de intensidade de cada *pixel*. À partir dos gradientes, é criado um histograma de gradientes para cada vídeo e este é comparado com os outros histogramas para a classificação da base.

Para testar um descritor global na base de vídeo KTH, Laptev *et al* [12] aplica o descritor de Zelnik [11] nessa base de duas maneiras diferentes: utilizando o múltiplas escalas temporais como o original e utilizando escalas temporais e espaciais. Ambos descritores apresentaram uma taxa de reconhecimento abaixo da encontrada pelos descritores locais testados no artigo, sendo que a segunda versão do descritor, quando aplicada à classificação da base de vídeos KTH utilizando um classificador SVM, alcançou uma precisão de aproximadamente 71%.

2.2.1 Análise do movimento utilizando tensores

Os tensores são robustas e poderosas ferramentas matemáticas que vem sendo exploradas com mais frequência na última década nas mais diversas aplicações. O tensor já foi largamente utilizado para o cálculo do fluxo óptico, porém no campo da análise do movimento e indexação de vídeo, as pesquisas ainda são escassas. Pesquisas utilizando tensores podem ser divididas em dois tipos: as que não os usam como descritor, mas sim, usam operações tensoriais no auxílio da análise e reconhecimento, como pode ser visto em [27] e [28]; e as que fazem uso das propriedades do tensor utilizando, assim, o tensor como descritor, como visto em [29] e [30].

É interessante notar que métodos de análise baseados em tensores geralmente não fazem uso do fluxo óptico como característica da sequência de imagens.

2.2.1.1 Abordagens utilizando operações tensoriais

Em [27] é introduzido um novo método para o reconhecimento de gestos e ações chamado Análise Canônica de Correlação Tensorial (do inglês - *Tensor Canonical Correlation Analysis*) que é uma extensão da clássica Análise Canônica de Correlação (CCA) para vetores multidimensionais. Esse método extrai características de correlação entre dois vídeos de maneira flexível e descritiva. Para a classificação, tanto da base de vídeos KTH quanto da base de gestos, é utilizado um classificador de vizinhos mais próximos

(do inglês - *Nearest Neighbor*). Esta classificação encontra uma precisão de 95,33% para a base de vídeo KTH.

Para Krausz e Bauckhage [28], uma ação é considerada como uma sequência de posições do corpo as quais são consideradas combinações lineares de partes do corpo. O processo é baseado em uma fatoração tensorial não-negativa que extrai imagens base que representam partes do corpo. Os pesos dos coeficientes são obtidos filtrando-se um quadro com esse conjunto de imagens base. Como essas imagens são obtidas através de uma fatoração tensorial não-negativa, elas são separáveis e podem ser eficientemente implementadas. Esse método é aplicado no reconhecimento de ações da base de vídeos Weizmann. Para reconhecer uma ação, cada quadro de um vídeo é filtrado com cada uma dessas imagens base e formam-se vetores de características. Para cada um desses vetores é determinado o conjunto de vizinhos mais próximos. O resultado final é obtido usando um método de votação.

2.2.1.2 Abordagens utilizando o tensor como descritor

Através do uso das características espectrais do tensor, Jia *et al* [29] propõe um método para reconhecimento de ações baseado em características em multiresolução. Uma série de silhuetas são transformadas em uma imagem denominada *Serials-Frame* da qual são extraídas características para a construção do espaço de autovetores e autovalores de um tensor chamado *Serials-Frame Tensor*. Um análise desse espaço é aplicada para separar as informações necessárias para o reconhecimento de diferentes ações. A base de vídeo usada para testar esse descritor foi criada pelos próprios autores.

Também utilizando a idéia de silhuetas, Khadem e Rajan [30] criam um conjunto das mesmas para formar um tensor de terceira ordem. Desta forma, o tensor formado compreende três modos: *pixels*, ações e pessoas. O tensor é projetado em diferentes espaços correspondendo a estes modos a partir da decomposição em valores singulares n-modal, a qual encontra as bases e os coeficientes do espaço de ações. Para uma dada sequência, os coeficientes resultantes são comparados com os coeficientes aprendidos para encontrar a classe da ação. Este método se mostrou mais eficiente que o clássico método de análise de componentes principais na classificação da base de vídeos Weizmann [23].

3 FUNDAMENTOS

Este capítulo trata dos fundamentos mais importantes para a compreensão deste trabalho. São eles:

- Fluxo óptico: uma forma de se extrair o movimento de uma sequência de imagens é através do cálculo do fluxo óptico. Na Seção 3.1 são apresentados o fluxo óptico e métodos para o seu cálculo.

- Modelagem de campos de deslocamento por base de polinômios: na Seção 3.2 é mostrada uma abordagem para se aproximar o fluxo óptico. A partir da modelagem de campos de deslocamento por base de polinômios é possível obter uma expressão polinomial do fluxo óptico, diminuindo assim a quantidade de informação necessária para a representação do movimento.

- Tensor de orientação: algumas propriedades e a interpretação geométrica deste tipo de tensor são apresentadas na Seção 3.3.

- Descritor global de movimento: na Seção 3.4 são mostradas algumas propriedades de descritores e apresentados dois descritores globais presentes na literatura, os descritores de Zelnik *et al* [11] e de Laptev *et al* [12].

3.1 Fluxo Óptico

A estimativa do movimento consiste em medir a projeção 2D no plano da imagem de um movimento real 3D. O movimento 2D é também chamado fluxo óptico e pode ser definido como o campo de velocidades descrevendo o movimento aparente das intensidades da imagem sob a hipótese de conservação da luminância. O fluxo óptico também pode ser chamado de campo de deslocamento.

A Figura 3.1 mostra um exemplo de cálculo do fluxo óptico entre dois quadros adjacentes. As duas imagens superiores representam dois quadros adjacentes de um vídeo do movimento de um passo realizado por um homem, a imagem inferior esquerda mostra o fluxo óptico calculado entre esses dois quadros. Finalmente, a última imagem representa

um detalhe do fluxo óptico no qual é apresentado o campo de direções do movimento da perna.

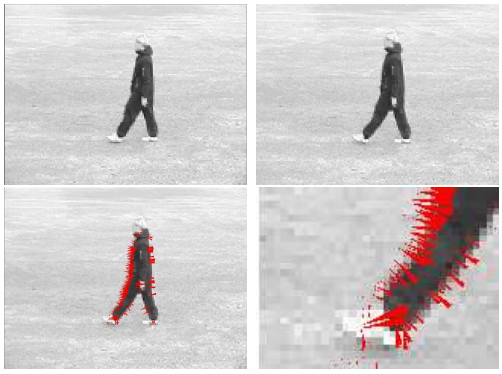


Figura 3.1: Exemplo do fluxo óptico calculado entre dois quadros adjacentes

Uma sequência de imagens pode ser representada por sua função de luminância $I(x_1, x_2, t)$. A hipótese de conservação da luminância significa que a luminância de um ponto físico da sequência de imagens não varia durante o tempo, isto é:

$$I(x_1, x_2, t) = I(x_1 + d_{x_1}, x_2 + d_{x_2}, t + 1) \tag{3.1}$$

Dado D o domínio espaço-temporal correspondendo à sequência $I(x_1, x_2, t) : D \to R$, Ω o domínio espacial de (x_1, x_2) e $v(x_1, x_2, t) : D \to \mathbb{R}^2$ o campo de velocidade instantânea no tempo t. O problema é encontrar v no tempo t. Levando em consideração a hipótese citada acima, tem-se o problema de Restrição do Fluxo Óptico (RFO) representado por:

$$\nabla I \ . \ \vec{v} + I_t \ = 0 \tag{3.2}$$

onde $\vec{v} = (v_{x_1}, v_{x_2}, 1)^T$

Esta equação só permite obter a componente $v_\perp$ paralela à $\hat{n} = \frac{\nabla I}{\|\nabla I\|}$. Isto é conhecido como Problema de Abertura. Só é possível encontrar o movimento aparente de um ponto efetuando o cálculo em uma vizinhaça limitada deste ponto. Só se consegue calcular a componente do movimento na direção do gradiente (isto é, perpendicular ao contorno). Além disso, a estimativa é impossível no caso de $\nabla I = 0$.

Admitindo uma superfície em movimento (o retângulo) sendo olhada através de pequenas janelas (simbolizadas por círculos), tem-se três casos possíveis (Figura 3.2):

- se o gradiente de intensidade é nulo, o movimento não é percebido (Figura 3.2a).

- se o gradiente de intensidade na janela é orientado em uma única direção, o movimento é percebido como normal ao contorno (Figura 3.2b).

- a combinação de informações de diferentes orientações dos gradientes permite encontrar o movimento real do objeto (Figura 3.2c).

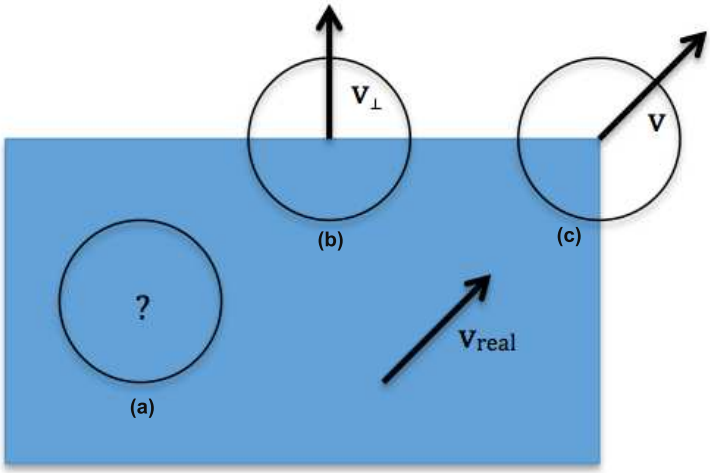


Figura 3.2: Ilustração do problema de abertura

Figura 3.3: Sinal do barbeiro

Um exemplo clássico da manifestação do problema de abertura é o "sinal do barbeiro"(tradução livre do francês - *enseigne du barbier*) mostrado na Figura 3.3, no qual tem-se a impressão que as linhas se deslocam para cima, enquanto, na verdade, elas se deslocam para a direita.

A estimativa local do movimento é apenas possível pela adição de restrições a partir da vizinhança espacial ou espaço-temporal; a solução obtida é então uma média do movimento nesta vizinhança. A mesma deve ser suficientemente grande para restringir a solução, sem cobrir as regiões contendo os diferentes movimentos. Assim, além da con-

servação da luminância, uma hipótese de continuidade espacial do fluxo óptico é necessária para determinar o movimento.

Os métodos para a computação do fluxo óptico podem ser divididos em quatro categorias [15]: os métodos por correlação, os métodos diferenciais, os métodos baseados em energia e os métodos baseados em fase. Alguns métodos diferenciais para a computação do fluxo óptico são: método de Lucas e Kanade [13], método de Horn e Shunck [14] e método de Augereau [1].

3.1.1 Método de Lucas e Kanade

Uma abordagem diferencial local para fazer a extração do fluxo óptico é o método de Lucas e Kanade [13]. Este método resolve a RFO a partir de médias quadradas de uma vizinhança em uma janela $W(x)$ (Equação 3.3). O fluxo v é obtido como o mínimo desta energia.

$$\int_{W(x)} (\nabla_{xt} I \ . \ \vec{v})^2 dx' = \int_{W(x)} \vec{v}^T (\nabla_{xt} I)(\nabla_{xt} I)^T \vec{v} dx' \tag{3.3}$$

No lugar de ter uma simples média na vizinhança x, pondera-se a RFO por uma função $h(x)$ que é normalmente um núcleo gaussiano de média nula e desvio padrão s_x (Equação 3.4):

$$h(x) = \frac{1}{2\pi s_x^2} e^{-\frac{x^T x}{2s_x^2}} \tag{3.4}$$

Assumindo

$$\langle f \rangle (x,t) = \int_{W(x)} h(x - x') f(x', t) dx' \tag{3.5}$$

e

$$S = \left\langle (\nabla_{x,t} I)(\nabla_{x,t} I)^T \right\rangle \tag{3.6}$$

que é a definição do tensor de estrutura, então o problema 3.3 pode ser reescrito como:

$$Argmin_{\vec{v}=(v_{x_1}, v_{x_2}, 1)} \vec{v}^T S \vec{v} \tag{3.7}$$

A minimização da energia é um problema de médias quadradas ponderadas e a solução se torna considerar o sistema de equações dado pela equação 3.8.

$$A \begin{pmatrix} v_{x_1} \\ v_{x_2} \end{pmatrix} = b \tag{3.8}$$

onde $A = \begin{pmatrix} \langle I_{x_1}^2 \rangle & \langle I_{x_1} I_{x_1} \rangle \\ \langle I_{x_1} I_{x_2} \rangle & \langle I_{x_2}^2 \rangle \end{pmatrix}$ e $b = \begin{pmatrix} \langle I_{x_1} I_t \rangle \\ \langle I_{x_2} I_t \rangle \end{pmatrix}$

Se o problema de abertura persistir, os autores recomendam o uso de uma janela de vizinhança maior.

3.1.2 Método de Horn e Schunck

Uma abordagem diferencial global para a extração do movimento é método de Horn e Schunck [14]. Para a resolução da RFO, Horn e Schunck adicionam uma restrição de coerência espacial do campo de deslocamentos. Eles procuram, então, o campo de deslocamento que resolve a equação 3.2 de tal forma que as derivadas espaciais $\nabla\tilde{v}$ sejam as mais fracas possíveis. Assim, deve-se minimizar a seguinte função:

$$\int\int_{\Omega} (\nabla I.\vec{v} + I_t)^2 + \lambda^2(||\nabla v_{x_1}||_2^2 + ||\nabla v_{x_2}||_2^2) d_{x_1} d_{x_2} \tag{3.9}$$

onde λ é um coeficiente que permite ponderar a influência da restrição de coerência.

O algoritmo proposto pelos autores é iterativo, de fácil implementação e converge para a solução. Para cada iteração n, o fluxo óptico é calculado da seguinte forma:

$$\begin{cases} v_{x_1}^{n+1} = \bar{v}_{x_1}^n - \frac{I_{x_1}(I_{x_1}\bar{v}_{x_1}^n + I_{x_2}\bar{v}_{x_2}^n + I_t)}{\alpha + I_{x_1}^2 + I_{x_2}^2} \\ v_{x_2}^{n+1} = \bar{v}_{x_2}^n - \frac{I_{x_2}(I_{x_1}\bar{v}_{x_1}^n + I_{x_2}\bar{v}_{x_2}^n + I_t)}{\alpha + I_{x_1}^2 + I_{x_2}^2} \end{cases} \tag{3.10}$$

onde $\bar{v}_{x_1}^n$ e $\bar{v}_{x_2}^n$ são respectivamente a média dos vizinhos de v_{x_1} e v_{x_2} na iteração n, I_{x_1} e I_{x_2} as derivadas espaciais em x_1 e x_2 de I e α uma constante real.

Esta nova restrição permite resolver o problema de abertura mas não permite obter campos de deslocamento que apresentam descontinuidades [14].

3.1.3 Método de Augereau

Um método pouco utilizado, porém muito interessante, é o método de Augereau [1]. Este método diferencial também utiliza o tensor de estrutura, como o método de Lucas e Kanade, mas utiliza as características espectrais do tensor. Este método extrai um fluxo óptico mais regular que o encontrado pelo método de Lucas e Kanade e por isso foi o escolhido para ser implementado neste trabalho.

Seja $\nabla I = (I_{x_1}, I_{x_2}, I_t)^T$ o gradiente de intensidades luminosas de um *pixel*. O tensor

de estrutura S deste *pixel* é então uma matriz simétrica definida positiva $S = \nabla I \nabla I^T$ dada por:

$$S = \nabla I \nabla I^T = \begin{pmatrix} I_{x_1}^2 & I_{x_1} I_{x_2} & I_{x_1} I_t \\ I_{x_1} I_{x_2} & I_{x_2}^2 & I_{x_2} I_t \\ I_{x_1} I_t & I_{x_2} I_t & I_t^2 \end{pmatrix} \tag{3.11}$$

Os elementos espectrais do tensor S são os seguintes autovalores β_i:

$$\beta_1^{(S)} = I_{x_1}^2 + I_{x_2}^2 + I_t^2, \quad \beta_2^{(S)} = \beta_3^{(S)} = 0 \tag{3.12}$$

Levando em consideração os autovetores de S, o autovetor associado à $\beta_1^{(S)}$ é o vetor gradiente $v_1^{(S)} = \nabla I$. Além disso, o subespaço gerado pelos dois outros autovetores $v_2^{(S)}$ e $v_3^{(S)}$, associados aos autovalores nulos, é ortogonal à ∇I. Assim, qualquer vetor pertencente ao núcleo de S é uma possível solução para a hipótese de iluminação constante. De fato, $v_2^{(S)}$ e $v_3^{(S)}$ podem ser escolhidos tal que ($v_1^{(S)}$, $v_2^{(S)}$, $v_3^{(S)}$) forme uma base ortogonal:

$$v_1^{(S)} = \begin{pmatrix} I_{x_1} \\ I_{x_2} \\ I_t \end{pmatrix}, \quad v_2^{(S)} = \begin{pmatrix} I_{x_2} \\ -I_{x_1} \\ 0 \end{pmatrix}, \quad v_3^{(S)} = \begin{pmatrix} I_{x_1} I_t \\ I_{x_2} I_t \\ -[I_{x_1}^2 + I_{x_2}^2] \end{pmatrix} \tag{3.13}$$

Uma abordagem local representativa do fluxo óptico pode ser obtida através do terceiro autovetor $v_3^{(S)}$. Para o caso de vídeos multicanais, existe uma extensão deste método que pode ser feita de modo simples na qual o fluxo óptico encontrado é uma combinação dos fluxos de cada canal.

3.2 Modelagem de Campos de Deslocamento por Base de Polinômios

Segundo Druon [9], pode-se definir um campo de deslocamento (ou fluxo óptico) F da seguinte forma:

$$F : \begin{array}{l} \Omega \subset \mathbb{R}^2 \to \mathbb{R}^2 \\ (x_1, x_2) \mapsto (V^1(x_1, x_2), V^2(x_1, x_2)) \end{array} \tag{3.14}$$

com $V^1 : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ e $V^2 : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ duas aplicações correspondendo respectivamente ao deslocamento horizontal e vertical dos pontos de coordenadas $(x_1, x_2) \in \Omega$.

Deseja-se aproximar funções reais de duas variáveis por funções polinomiais. Utiliza-se então polinômios definidos em $\mathbb{R}[x_1, x_2]$ da seguinte forma:

$$P_{K,L} = \sum_{k=0}^{K} \sum_{l=0}^{L} c_{k,l} (x_1)^k (x_2)^l \tag{3.15}$$

onde $K \in \mathbb{N}$ é o grau máximo de x_1, $L \in \mathbb{N}^*$ é o grau máximo de x_2 e $c_{k,l}$ é o conjunto de coeficientes reais do polinômio. O grau do polinômio é então $K + L$ [9].

3.2.1 Geração de uma base ortogonal bidimensional

Pode-se gerar uma família de funções ortogonais a partir da fórmula de recorrência à três termos:

$$\begin{cases} P_{-1,j} = 0 \\ P_{i,-1} = 0 \\ P_{0,0} = 1 \\ P_{i+1,j} = (a_i x_1 + b_i) P_{i,j} - c_i P_{i-1,j} \\ P_{i,j+1} = (a_j x_1 + b_j) P_{i,j} - c_j P_{i,j-1} \end{cases} \tag{3.16}$$

Esta fórmula representa a ortogonalização de Gram-Schmidt.

A partir dos valores de a_n, b_n e c_n, pode-se gerar diferentes famílias de polinômios ortogonais conhecidos apresentados na Tabela 3.1. Estes polinômios são ortogonais dois à dois no domínio Ω em relação ao produto escalar definido por 3.17 relativo à função de pesos $\omega(x_1, x_2)$.

$$\langle f_1(x) | f_2(x) \rangle = \int_{\Omega} f_1 f_2 \omega(x) dx \tag{3.17}$$

Família	Ω	$w(x_1, x_2)$	a_n	b_n	c_n
Legendre	$[-1;1]^2$	1	$\frac{2n+1}{n+1}$	0	$\frac{n}{n+1}$
Tchebychev 1	$[-1;1]^2$	$\prod_{i=1}^{2}(1-x_i^2)^(-1/2)$	2	0	1
Tchebychev 2	$[-1;1]^2$	$\prod_{i=1}^{2}(1-x_i^2)^(1/2)$	2	0	1
Laguerre	$[0;\infty]^2$	$exp^{-(x_1+x_2)}$	$\frac{-1}{n+1}$	$\frac{2n+1}{n+1}$	$\frac{n}{n+1}$
Hermite	$[-\infty;\infty]^2$	$exp^{\frac{-(x_1^2+x_2^2)}{2}}$	2	0	2n

Tabela 3.1: Parâmetros utilizados na fórmula de recorrência 3.16 para a geração de algumas famílias de polinômios ortogonais conhecidos [9]

.

A base bidimensional $B = \{P_{i,j}\}$ é então composta de polinômios ortonormais. Nota-

se g o grau mais elevado dos polinômios que compõe a base. Uma base bidimensional de grau g é constituída pelo conjunto de polinômios $P_{i,j}$ com $i + j \leq g$:

$$B_g = \{P_{0,0}, P_{0,1}, ..., P_{0,g}, P_{1,0}, P_{1,g-1}, ..., P_{g-1,0}, P_{g-1,1}, P_{g,0}\} \tag{3.18}$$

Assim, o número de polinômios que compõe uma base de grau g é :

$$n_g = \frac{(g+1)(g+2)}{2} \tag{3.19}$$

Uma representação tabular de uma base de polinômios B_g é dada na Figura 3.4.

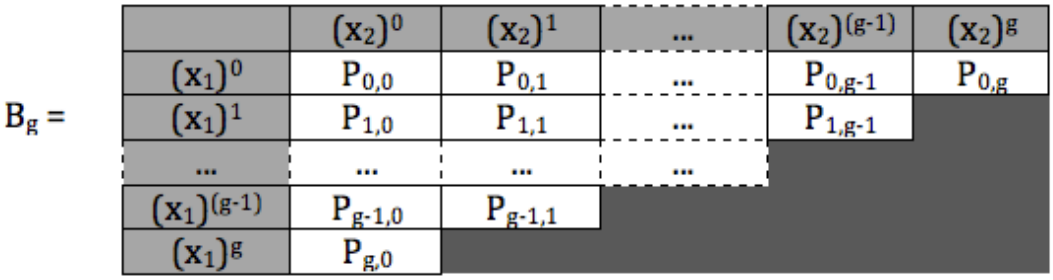


Figura 3.4: Representação tabular de uma base bidimensional de grau g.

3.2.2 Aproximação de um campo de deslocamento

Para expressar o campo de deslocamento F por combinações lineares de diferentes polinômios $P_{i,j}$ da base ortonormal B_g, a idéia de Druon [9] é projetar as funções $V^1(x_1, x_2)$ e $V^2(x_1, x_2)$ sobre cada polinômio $P_{i,j}$ da base. Pode-se expressar $F = (\tilde{V}^1(x_1, x_2), \tilde{V}^2(x_1, x_2))$ como sendo:

$$\begin{cases} \tilde{V}^1(x_1, x_2) = \sum_{i=0}^{g} \sum_{j=0}^{g-1} \tilde{v}^1_{i,j} P_{i,j} \quad com \quad \tilde{v}^1_{i,j} = \langle V^1(x_1, x_2) \mid P_{i,j} \rangle \\ \\ \tilde{V}^2(x_1, x_2) = \sum_{i=0}^{g} \sum_{j=0}^{g-1} \tilde{v}^2_{i,j} P_{i,j} \quad com \quad \tilde{v}^2_{i,j} = \langle V^2(x_1, x_2) \mid P_{i,j} \rangle \end{cases} \tag{3.20}$$

Os coeficientes $\tilde{v}^1_{i,j}$ e $\tilde{v}^2_{i,j}$ são calculados a partir do produto escalar:

$$\begin{cases} \tilde{v}^1_{i,j} = \int \int_{\Omega} V^1(x_1, x_2) P_{i,j} \omega(x_1, x_2) dx_1 dx_2 \\ \\ \tilde{v}^2_{i,j} = \int \int_{\Omega} V^2(x_1, x_2) P_{i,j} \omega(x_1, x_2) dx_1 dx_2 \end{cases} \tag{3.21}$$

com $i + j \leq g$.

Conhecendo os coeficientes de projeção $\tilde{v}^1_{i,j}$ e $\tilde{v}^2_{i,j}$, pode-se determinar a expressão analítica do modelo através da equação 3.20. Finalmente, a partir desta expressão analítica, pode-se reconstruir a aproximação do campo original analisando os polinômios $\tilde{V}^1$ e $\tilde{V}^2$ em todos os pontos desejados. Quanto maior o grau da base, melhor é a aproximação do campo de deslocamento.

O processo de modelagem e reconstrução é dado pela Figura 3.5. Este processo pode ser dividido em três partes principais: a fase de projeção de um campo em uma base permitindo calcular os coeficientes de projeção; a fase de análise que permite determinar a expressão analítica do movimento a partir dos coeficientes de projeção e da base; e a fase de reconstrução na qual é calculado o campo de deslocamento modelado a partir da expressão analítica, ou seja, é calculado o fluxo óptico aproximado.

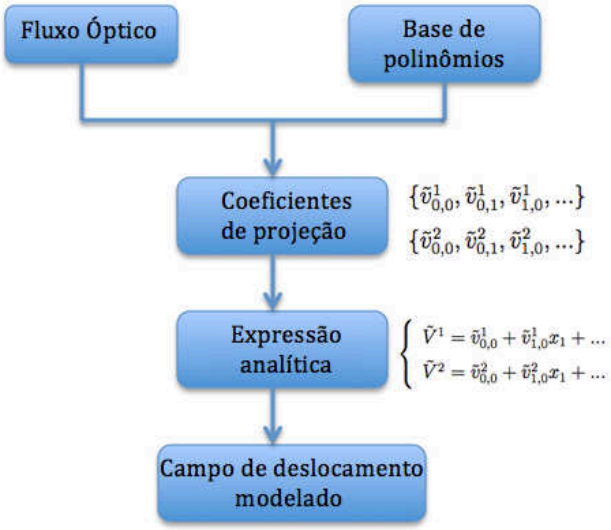


Figura 3.5: Processo de modelagem e reconstrução de um campo de deslocamento.

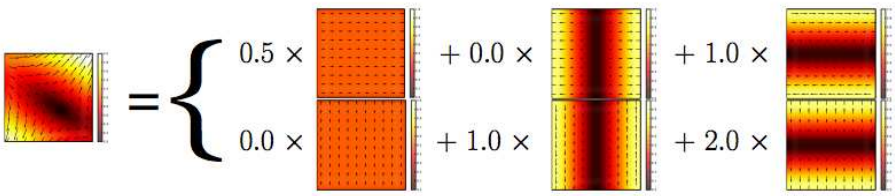


Figura 3.6: Exemplo de combinações lineares de polinômios permitindo modelar um campo sintético [9]

Um exemplo de combinações lineares de polinômios permitindo modelar um campo sintético é apresentado na Figura 3.6.

Neste trabalho utiliza-se os polinômios de Legendre. A partir dos parâmetros dados na Tabela 3.1, a fórmula de recorrência 3.16 se torna:

$$\begin{cases} P_{-1,j} = 0 \\ P_{i,-1} = 0 \\ P_{0,0} = 1 \\ P_{i+1,j} = \frac{2i+1}{i+1} x_1 P_{i,j} - \frac{i}{i+1} P_{i-1,j} \\ P_{i,j+1} = \frac{2j+1}{j+1} x_2 P_{i,j} - \frac{j}{j+1} P_{i,j-1} \end{cases} \tag{3.22}$$

A escolha dos polinômios de Legendre se justifica por diversas razões. A principal delas é que deseja-se efetuar uma modelagem global dos campos de deslocamento. A função de pesos $\omega(x_1, x_2) = 1$ permite dar a mesma importância a todos os vetores do campo. Além disso, esta função de pesos simplifica consideravelmente as equações que permitem calcular os coeficientes de projeção, o que diminui o tempo de cálculo [9].

3.3 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.

Pela definição de [31], o tensor de orientação é uma matriz real e simétrica, desta forma pode ser decomposta utilizando o teorema espectral da seguinte forma:

$$T = \sum_{i=1}^{n} \lambda_i T_i \tag{3.23}$$

onde λ_i são os autovalores do tensor T.

Projetando T_i sobre um espaço de dimensão m, tem-se a seguinte decomposição:

$$T_i = \sum_{s=1}^{m} e_s e_s^T \tag{3.24}$$

onde $\{e_1,...,e_m\}$ é uma base de R^m. Uma decomposição interessante do tensor de ori-

entação T é dada por

$$T = \lambda_n T_n + \sum_{i=1}^{n-1} (\lambda_i - \lambda_{i+1}) T_i \tag{3.25}$$

onde λ_i são os autovalores correspondentes à cada autovetor e_i. Esta decomposição é interessante por causa de sua interpretação geométrica. De fato, em $\mathbb{R}^3$, o tensor de orientação T decomposto utilizando a equação 3.25 pode ser representado por uma lança, um prato e uma esfera (Figura 3.7).

$$T = (\lambda_1 - \lambda_2) T_1 + (\lambda_2 - \lambda_3) T_2 + \lambda_3 T_3. \tag{3.26}$$

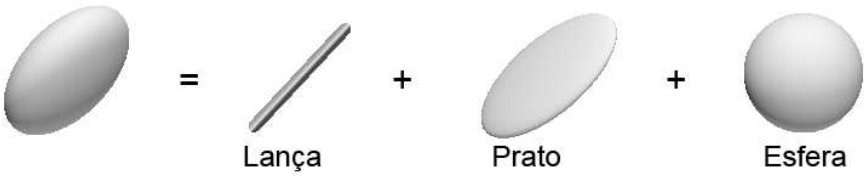


Figura 3.7: Decomposição de um tensor em $\mathbb{R}^3$ em suas componentes lança, prato e esfera.

O tensor de $\mathbb{R}^3$ decomposto como na equação 3.26, com os autovalores $\lambda_1 \geq \lambda_2 \geq \lambda_3$, pode ser interpretado como:

- $\lambda_1 >> \lambda_2 \approx \lambda_3$ corresponde aproximadamente ao tensor linear, com a componente lança dominante.
- $\lambda_1 \approx \lambda_2 >> \lambda_3$ corresponde aproximadamente ao tensor prato, com a componente prato dominante.
- $\lambda_1 \approx \lambda_2 \approx \lambda_3$ corresponde aproximadamente ao tensor isotrópico, com a componente esfera dominante e nenhuma orientação dominante.

Pode-se criar um tensor de orientação a partir de um vetor utilizando seu transposto da seguinte forma:

$$T = \vec{v}\vec{v}^T \tag{3.27}$$

onde T é o tensor criado a partir do produto entre o vetor $\vec{v}$ e seu transposto $\vec{v}^T$. Então, o tensor T é uma matriz $n \times n$, com n sendo o tamanho do vetor $\vec{v}$. Deste modo, é possível

considerar o tensor de orientação como sendo uma medida de covariância dos elementos do vetor.

3.4 Descritor Global de Movimento

Nesta seção serão apresentados os descritores globais de Zelnik [11] e de Laptev [12]. Foi necessário implementar estes descritores para que eles fossem avaliados nas mesmas condições do descritor proposto.

3.4.1 Algumas propriedades de descritores

Um descritor é 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 [32].

Os descritores de vídeos e imagens podem ser divididos em duas grandes categorias: descritores locais, que visam focar em certos pontos ou áreas da imagem; e descritores globais, que visam descrever todo o conteúdo da imagem. O uso de cada um destes tipos de descritores depende da aplicação, por exemplo, no caso do reconhecimento de ações nota-se uma maior preferência pelos descritores locais aos globais, uma vez que, o primeiro alcança precisões mais altas que as alcançadas pelo segundo. Isso não impede o uso de descritores globais e não significa que estes também não alcancem boas precisões.

Os descritores geralmente utilizados em imagens são baseados em cor, textura e forma. Estas informações podem ser utilizadas individualmente ou combinadas. A cor e a textura podem ser utilizadas tanto de maneira global quanto para partes da imagem, dividindo-a em regiões de interesse para a extração de características locais. A combinação de cor e textura pode ser ainda utilizada para detecção de formas na imagem [33].

Em se tratando de sequências de imagens, uma das características mais utilizadas é o movimento. Este constitui uma fonte de informação visual de extrema importância e oferece informações sobre a estrutura tridimensional da cena, a trajetória de um objeto e a atividade que está acontecendo.

Alguns exemplos de descritores de vídeo são: SIFT, que é um detector de pontos de interesse e descritor local que procura por pontos que apresentam invariâcia em relação à posição, escala e localização; e o histograma de gradientes, um descritor global que conta o número de ocorrências das orientações dos gradientes presentes em uma imagem.

Outra forma de se montar um descritor é o uso de um pacote de características (do Inglês - *bag-of-features*). Assim, cada vídeo ou imagem é representado por um conjunto de descritores.

3.4.2 Os descritores de Zelnik e de Laptev

Zelnik *et al* [11] apresentam um descritor global de movimento baseado em histogramas de gradientes. A idéia do método é extrair características em multiescala de um vídeo e construir uma distribuição empírica associada a ação que acontece na cena. Segundo os autores, duas ações serão consideradas similares caso elas possam gerar o mesmo processo estocástico, isto é, se suas distribuições empíricas, em uma mesma escala, são similares. No caso, a distribuição empírica escolhida é o histograma de gradientes.

Primeiramente é construída uma pirâmide temporal de todo o vídeo a partir de uma filtragem passa-baixa e de uma amostragem da sequência de imagens na direção do tempo. A pirâmide temporal de uma sequência S é então uma pirâmide de sequências $S^1(= S), S^2, ..., S^L$, na qual os quadros das sequências possuem a mesma dimensão e cada sequência S^l possui a metade do número de quadros da sequência de resolução mais alta S^{l-1}. Geralmente, são utilizadas 3 ou 4 escalas temporais.

Para cada sequência da escala l da pirâmide temporal, é estimado o gradiente de intensidade $(S^l_{x_1}, S^l_{x_2}, S^l_t)$ de cada ponto (x_1, x_2, t). Assim, serão calculados os gradientes de intensidade de cada quadro de cada sequência da escala l. A partir destes, constrói-se um histograma de gradientes para cada escala, isto é, conta-se o número de ocorrências das orientações dos gradientes de cada escala. São utilizados 256 elementos para cada histograma de cada escala, desta forma, o tamanho do descritor é $L \cdot k \cdot 256$, onde L é o número de escalas e k o número de dimensões (no caso de vídeos tem-se 3 dimensões - x, y e t).

A distância utilizada para medir a similaridade entre os histogramas é a distância χ^2 dada por:

$$\chi^2(x, y) = \sum_i \frac{(x(i) - y(i))^2}{x(i) + y(i)} \tag{3.28}$$

De acordo com Zelnik *et al* [11], a avaliação deste descritor é feita através de um método de agrupamento (*clustering*) de ações de vídeos da base Weizmann.

Em Laptev *et al* [12] é apresentado uma comparação entre diversos descritores locais e dois descritores globais. O primeiro deles é o descritor de Zelnik *et al* utilizando 3 escalas temporais. O segundo deles é uma extensão do descritor de Zelnik *et al* no qual é levado em consideração os gradientes de múltiplas escalas espaciais. No caso, são utilizadas 3 escalas temporais e 3 espaciais.

As escalas espaciais formam uma pirâmide de sequências $S_1^1, S_1^2, ..., S_1^L, S_2^1, ..., S_P^L$, cujas sequências de uma mesma escala espacial tem quadros de mesma dimensão e metade do número de quadros da sequência de resolução temporal mais alta. Já as sequências de escalas espaciais diferentes possuem o mesmo número de quadros da sequência de resolução espacial mais alta, porém com metade da dimensão.

Para reconhecer as ações da base de vídeo KTH [10] usando os dois tipos de descritores globais, foi feita a classificação desta base através do método de vizinhos mais próximos. No caso do descritor estendido (com escalas temporais e espaciais), a classificação também foi feita utilizando um classificador SVM de núcleo χ^2.

4 PROPOSIÇÃO DE UM DESCRITOR TENSORIAL

Neste capítulo apresenta-se o descritor global de movimento proposto neste trabalho. Este descritor é baseado no tensor de orientação formado à partir dos coeficientes dos polinômios de Legendre calculados para cada quadro de um vídeo. Os coeficientes são encontrados através da projeção do fluxo óptico nos polinômios de Legendre, obtendo-se uma representação polinomial do movimento.

Na Seção 4.1, apresenta-se este descritor e um exemplo de sua aplicação na classificação de vídeos. O uso de polinômios de Legendre que representam o fluxo óptico, sua codificação na forma de tensores e a proposta de acumulá-los para a criação do descritor é a principal contribuição deste trabalho.

Afim de melhorar o desempenho do descritor em aplicações de classificação, são propostas as seguintes variações:

- Inserindo coerência temporal: Nesta variação adiciona-se a derivada de primeira ordem na criação do tensor, como consequência o descritor aumenta de tamanho, porém se torna capaz de capturar variações no tempo. Esta variação é apresentada na Seção 4.2.

- Agrupamento de tensores no tempo: Nesta variação explora-se a idéia de pacote de características. O descritor é formado por um conjunto de tensores calculados entre intervalos de quadros do vídeo, assim cada um destes tensores representará o movimento médio de uma porção menor do vídeo. Esta variação é apresentada na Seção 4.3.

- Agrupamento de tensores no espaço: Esta variação também explora a idéia de pacote de características. O descritor é formado pelo conjunto de tensores calculados a partir de divisões dos quadros do vídeo, assim cada um destes tensores representa uma porção do movimento presente na cena. Esta variação é apresentada na Seção 4.4.

- Descritor obtido em janela deslizante: Nesta variação, no lugar de se aproximar todo o fluxo óptico presente nos quadros, aproxima-se uma região de interesse. Desta maneira, apenas o campo de deslocamentos da região com os movimentos mais representativos do vídeo é usado para calcular o descritor. Esta variação é apresentada na Seção 4.5.

A avaliação e comparação de todos estes descritores é encontrada no Capítulo 5.

4.1 Descritor tensorial baseado em polinômios de Legendre que representam o fluxo óptico

A proposta deste trabalho é criar um novo descritor global de movimento baseado no tensor de orientação. Este tensor é criado a partir dos coeficientes $\tilde{v}^1_{i,j}$ e $\tilde{v}^2_{i,j}$ (Eq. 3.21) encontrados pela modelagem do fluxo óptico por base de polinômios.

A partir da aproximação do campo de deslocamentos é criado um vetor $\vec{v}_f$ para cada quadro f do vídeo:

$$\vec{v}_f = [\tilde{v}^1_{0,0}, ..., \tilde{v}^1_{g,0}, \tilde{v}^2_{0,0}, ..., \tilde{v}^2_{g,0}] \in \mathbb{R}^m \tag{4.1}$$

onde $\tilde{v}^1_{i,j}$ e $\tilde{v}^2_{i,j}$ são os coeficientes encontrados a partir de uma projeção do campo de deslocamentos do quadro f em uma base de grau g.

Em seguida, utilizando o vetor de coeficientes $\vec{v}_f$, cria-se um tensor de orientação para cada quadro f de acordo com a Equação 3.27:

$$T_f = \vec{v}_f \ \vec{v}_f^T \tag{4.2}$$

Desta forma, o tensor de orientação T_f captura a informação de covariância entre os coeficientes $\tilde{v}^1_{i,j}$ e $\tilde{v}^2_{i,j}$. O tensor T_f é uma matriz $2 \cdot n_g \times 2 \cdot n_g$, com $2 \cdot n_g$ sendo o tamanho do vetor de coeficientes. A Figura 4.1 mostra um exemplo de como um tensor é formado a partir dos coeficientes de uma base de grau 1. Para esta base, tem-se 6 coeficientes para cada quadro do vídeo (3 para o polinômio em x_1 e 3 para o polinômio em x_2), assim o tensor formado é uma matriz 6×6.

Esse tensor (Eq. 4.2) é classificado como do tipo lança e carrega informação do polinômio do quadro f m-dimensional. Este tipo de tensor captura a informação direcional do vetor de coeficientes $\vec{v}_f$ e não captura as incertezas do movimento. Então, do ponto

de vista de classificação, isoladamente este tensor não é um bom representante do movimento. Porém combinando esses tensores utilizando uma soma simples como em [31], existe a possibilidade de haver uma variação dos coeficientes dos polimônios e, assim, o tensor formado por vários quadros tende a capturar a variação do movimento da cena.

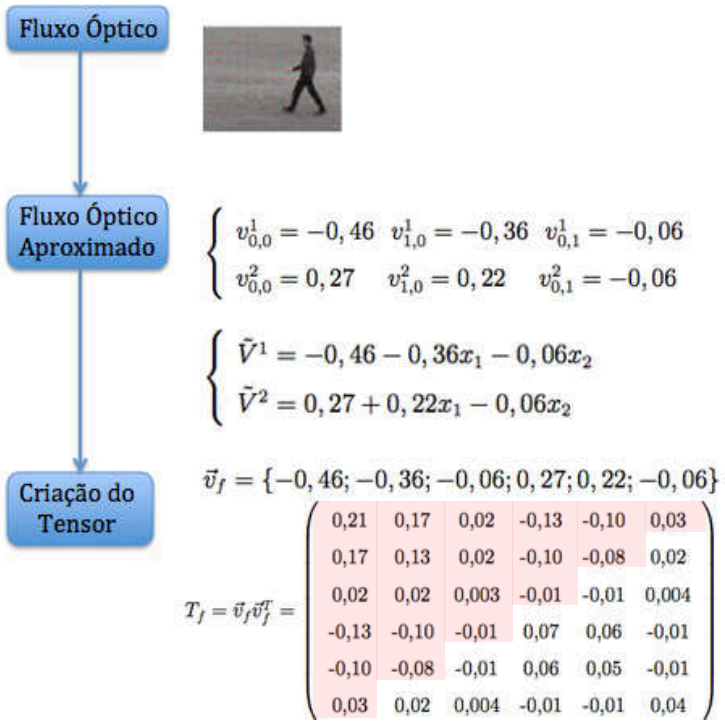


Figura 4.1: Criação de um tensor T a partir dos coeficientes $\tilde{v}^1_{i,j}$ e $\tilde{v}^2_{i,j}$

Desta forma, para expressar o movimento médio de diversos quadros consecutivos, o descritor é formado pela soma dos tensores dos quadros de um vídeo:

$$\bar{D} = \sum_{f=a}^{b} T_f \tag{4.3}$$

onde $[a, b]$ é o intervalo dos quadros que contém o movimento mais representativo do vídeo. Assim, é feita uma combinação m-dimensional que captura a dispersão dos tensores neste espaço. É possível utilizar somente um trecho do vídeo para compor o descritor, isto é, pode-se escolher a parte mais representativa do movimento.

O espaço de autovetores e autovalores do tensor final captura informações importantes como, por exemplo, as anisotropias que devem ser exploradas. Seja $\{\lambda_1, \lambda_2, \cdots, \lambda_m\}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, os valores próprios de um tensor T. Quanto maior o valor de $\lambda_1 - \lambda_2$, melhor o tensor expressa o movimento médio dado pelo autovetor principal. Ou seja, o

tensor é um bom representante do movimento acumulado à medida que é anisotrópico. Caso seja isotrópico, o tensor não consegue representar o movimento acumulado.

Uma medida da anisotropia α_T do tensor T em $\mathbb{R}^3$ é dada por:

$$\alpha_T = \sqrt{\frac{(\lambda_1 + \lambda_2)^2 + (\lambda_2 + \lambda_3)^2}{2(\lambda_1 + \lambda_2 + \lambda_3)^2}} \tag{4.4}$$

onde λ_1, λ_2 e λ_3 são os autovalores do tensor T. Quanto mais próxima de zero é a anisotropia α_T, mais isotrópico é o tensor T.

A anisotropia α_T é utilizada para calcular o quão representativo é o tensor calculado para o quadro T. No caso, quanto mais próxima de 1, mais o tensor é anisotrópico e assim não deve entrar na criação do descritor (Equação 4.3). Porém, testes feitos com o descritor criado desta forma não apresentaram diferenças significantes nos resultados, assim, não utiliza-se a anisotropia para a criação do descritor.

Devido a essas propriedades do tensor de orientação, o descritor proposto possui certas limitações. Quando o movimento da cena possui descontinuidades, o tensor se tornará isotrópico, ou seja, não terá uma direção principal, o que não é interessante para a descrição do movimento. Um exemplo de descontinuidade é quando cada quadro do vídeo contém o movimento em uma direção diferente, como por exemplo, tem-se um quadro com movimento para cima, em seguida um quadro com movimento para baixo, outro com movimento para a esquerda e por fim um para a direita. É necessário, portanto, que o movimento presente na cena seja suave e não tenha mudanças bruscas.

Outra limitação do descritor proposto é o tamanho do vídeo a ser descrito. Quanto maior o número de quadros, melhor será a média do movimento, porém, o tensor tende a se tornar isotrópico. Assim, este descritor não é interessante para descrever vídeos muito longos, em média, vídeos com mais de 2000 quadros (1 minuto e 20 segundos para o caso de vídeos com 25 quadros por segundo).

Como tensores podem ser combinados linearmente, a função de distância associada a este descritor pode ser definida utilizando a norma l_2. Logo, dado dois tensores A e B, pode-se definir a distância entre eles da seguinte forma:

$$||A - B|| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} |A_{ij} - B_{ij}|^2} \tag{4.5}$$

É interessante destacar que a norma l_1 também foi testada, porém não apresentou bons resultados.

4.1.1 Exemplo de aplicação do descritor para a classificação de vídeos

Para cada vídeo é calculado o fluxo óptico utilizando o método de Augereau [1]. Este método é utilizado por apresentar campos de descolamentos mais regulares que os encontrados pelo método de Lucas e Kanade.

A partir destes campos de deslocamento, é feita a modelagem por base de polinômios de Legendre e encontra-se n_g coeficientes para o polinômio em x_1 e para o polinômio em x_2. Assim, tem-se $2 \cdot n_g$ coeficientes para cada quadro do vídeo. Com estes coeficientes é criado o vetor $\vec{v}_f$ (Eq. 4.1) para cada quadro o qual é transformado em um tensor de orientação, como mostra a Equação 4.2.

Devido à sua natureza simétrica, não é necessário utilizar todos os coeficientes da matriz no descritor, ele se torna então uma matriz triangular com $n_g \cdot (2 \cdot n_g + 1)$ elementos.

Por fim, o descritor é composto pela soma dos tensores dos quadros mais representativos do vídeo. Para que se possa comparar os descritores de diferentes vídeos, é necessário normalizá-los pela norma l_2, já que não é possível saber quantos quadros foram utilizados para os compor. A comparação entre os descritores pode ser feita utilizando a distância l_2 (Eq. 4.5). Assim, quanto menor a distância entre dois vídeos, mais similares eles são.

Para utilizar este descritor para o reconhecimento de ações de uma base de vídeos, é calculado o descritor para cada vídeo pertencente à base. Por conseguinte, é aplicado um método de classificação que indicará quais vídeos pertencem a quais tipos de ação. Para que a classificação atinja uma boa precisão, isto é, consiga classificar um bom número de vídeos corretamente, é necessário que a função de distância utilizada seja a l_2.

A Figura 4.2 apresenta um exemplo da aplicação do descritor para a classificação de vídeos. Esta aplicação possui duas etapas principais: a criação do descritor (Figura 4.1) e a comparação entre eles. Na Figura 4.2 são mostrados três vídeos da base de vídeo KTH [10] que representam duas ações do tipo *walking* e uma do tipo *boxing*. O vídeo representado pela primeira imagem é comparado aos outros dois vídeos.

Para cada um dos vídeos é calculado o descritor $\bar{D}$ (Equação 4.3), somando-se os tensores de todos os quadros do vídeo. Para comparar os descritores é utilizado a distância

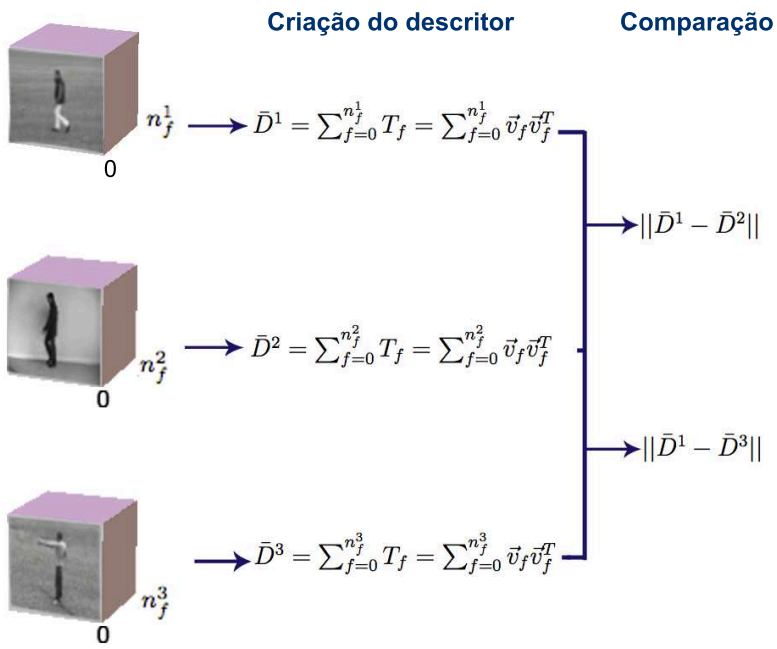


Figura 4.2: Exemplo de uso do descritor na classificação de vídeos

l_2 (Equação 4.5).

Todas as variações do descritor apresentadas a seguir podem ser utilizadas da mesma forma na classificação de vídeos.

4.2 Inserindo coerência temporal ao descritor

Com o objetivo de adicionar informação de velocidade ao descritor, foi incluído na criação do mesmo a derivada de primeira ordem em relação ao tempo do vetor de coeficientes $\vec{v}_f$ (Equação 4.1).

Dado o vetor de coeficientes $\vec{v}_f$, sua primeira derivada em relação ao tempo é dada por:

$$\partial_t \vec{v}_f = [\frac{\tilde{v}_{i,j}^1(f) - \tilde{v}_{i,j}^1(f-1)}{\Delta t}, \frac{\tilde{v}_{i,j}^2(f) - \tilde{v}_{i,j}^2(f-1)}{\Delta t}]_{i+j<g} \tag{4.6}$$

onde $\tilde{v}_{i,j}^1(f)$ e $\tilde{v}_{i,j}^2(f)$ representam os coeficientes calculados para o quadro f. Esta abordagem só é válida porque o tempo entre os quadros não varia, assim, pode-se considerar $\Delta t = 1$.

O novo vetor de coeficientes será então $\vec{v}_f^{novo} = \{\vec{v}_f, \partial_t \vec{v}_f\}$. A partir deste novo vetor é calculado o tensor de orientação para cada quadro f e o descritor é formado pela soma dos tensores dos quadros do vídeo, como explicado na Seção 4.1. Desta forma, o descritor

passa a ter $2 \cdot n_g \cdot (4 \cdot n_g + 1)$ coeficientes.

4.3 Agrupamento de tensores no tempo

Outra forma de se montar um descritor é através do uso de pacotes de características. Considerando o vídeo como um volume espaço-temporal, é possível montar um tensor para cada trecho do vídeo dividindo-o no tempo ou, ainda, no espaço. Então, o descritor será formado por um conjunto de tensores calculados para cada porção do vídeo.

Quando se tem pacotes de características formados por tensores construídos a partir da divisão no tempo, cada um destes tensores captura a informação do movimento médio entre o intervalo de quadros $[a, b]$ como visto da Equação 4.3. Desta forma, o tensor representará o movimento médio de uma porção menor do vídeo.

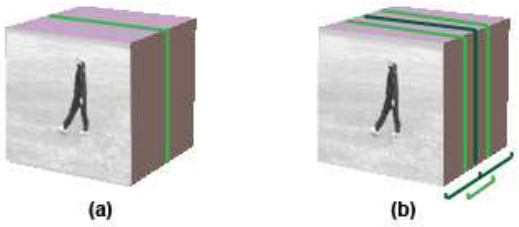


Figura 4.3: Tipos de divisão do vídeo no tempo.

A Figura 4.3 apresenta dois tipos de divisão no tempo. Sendo n_f o número total de quadros de um vídeo, a divisão presente na Figura 4.3(a) representa uma divisão de $[0, \frac{n_f}{2}]$ e de $[\frac{n_f}{2} + 1, n_f]$. Já a divisão presente na Figura 4.3(b) representa a divisão de $[0, \frac{n_f}{2}]$, de $[\frac{n_f}{2} + 1, n_f]$ e de $[\frac{n_f}{4}, \frac{3n_f}{4}]$. Para cada um desses intervalos será criado um tensor (Equação 4.3) e o descritor será o conjunto desses tensores.

A divisão de 4.3(b) mostra uma redundância na criação do descritor, isto é, alguns quadros são usados novamente para a criação do descritor. Esta redundância representa um reforço de um trecho da ação contida no vídeo, o qual é suposto ser mais representativo entre os quadros $[\frac{n_f}{4}, \frac{3n_f}{4}]$ da cena.

O uso de pacote de características com divisões no tempo não muda o processo de criação do tensor, apenas será alterado o intervalo de quadros para qual ele será calculado e o número de tensores que representa um vídeo.

4.4 Agrupamento de tensores no espaço

Quando o pacote de características é formado por tensores construídos a partir da divisão no espaço, cada um destes tensores representa uma porção do movimento presente na cena.

A Figura 4.4 apresenta três tipos de divisão no tempo. Sendo $n_{x_1} \times n_{x_2}$ a dimensão dos quadros do vídeo, a divisão presente na Figura 4.4 (a) representa uma divisão de $[0, n_{x_1}] \times [0, \frac{n_{x_2}}{3}]$,$[0, n_{x_1}] \times [\frac{n_{x_2}}{3} + 1, \frac{2n_{x_2}}{3}]$ e $[0, n_{x_1}] \times [\frac{2n_{x_2}}{3} + 1, n_{x_2}]$. Já na Figura 4.4 (b) a divisão é de $[0, \frac{n_{x_1}}{2}] \times [0, \frac{n_{x_2}}{2}]$, $[\frac{n_{x_1}}{2} + 1, n_{x_1}] \times [0, \frac{n_{x_2}}{2}]$, $[0, \frac{n_{x_1}}{2}] \times [\frac{n_{x_2}}{2} + 1, n_{x_2}]$ e $[\frac{n_{x_1}}{2} + 1, n_{x_1}] \times [\frac{n_{x_2}}{2} + 1, n_{x_2}]$. Na Figura 4.4 (c) está presente a mesma divisão da Figura 4.4 (b) mais uma divisão de $[\frac{n_{x_1}}{4}, \frac{3n_{x_1}}{4}] \times [\frac{n_{x_2}}{4}, \frac{3n_{x_2}}{4}]$. Da mesma forma como na divisão no tempo, para cada um desses intervalos será criado um tensor e o descritor será o conjunto desses tensores.

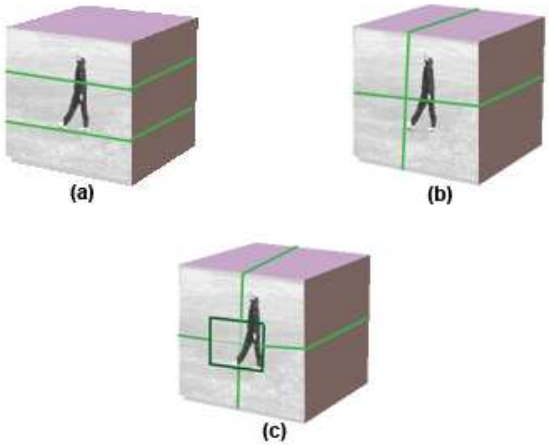


Figura 4.4: Tipos de divisão do vídeo no espaço.

De forma similar à divisão de 4.3(b), a divisão de 4.4(c) mostra uma redundância na criação do descritor. Neste caso, a redundância representa o reforço da região central dos quadros da cena, supondo ser esta a região de movimento mais representativo do vídeo.

O uso de pacote de características com divisões no espaço altera o cálculo dos coeficientes $\tilde{v}^1_{i,j}$ e $\tilde{v}^2_{i,j}$ e o número de tensores que representa um vídeo, mas não altera o processo de criação do tensor.

4.5 Descritor obtido em janela deslizante

No lugar de se aproximar todo o fluxo óptico contido em um quadro do vídeo, é possível aproximar apenas uma região de interesse (Figura 4.5). Desta maneira, apenas o campo de deslocamentos da região com os movimentos mais representativos do vídeo é usado para o cálculo dos coeficientes $\tilde{v}^1_{i,j}$ e $\tilde{v}^2_{i,j}$.

Para separar a região com os movimentos mais importantes é utilizada uma janela deslizante de tamanho fixo em torno da região de interesse. No caso do reconhecimento de ações, esta janela é colocada em volta do ser humano presente na sequência (Figura 4.5). Para calcular a posição da janela encontra-se o centróide do fluxo óptico, o qual vai sempre acompanhar o centro da região de movimento mais significativo. Logo, a janela será colocada em volta deste centróide.

Figura 4.5: Representação da região com movimentos mais representativos de um vídeo da base KTH.

Esta abordagem não muda o processo de criação do descritor, apenas altera o cálculo dos coeficientes $\tilde{v}^1_{i,j}$ e $\tilde{v}^2_{i,j}$. Assim, pode-se criar o tensor final tanto com derivada quanto sem.

5 RESULTADOS E ANÁLISE COMPARATIVA

Neste capítulo será apresentada a avaliação dos cinco descritores criados e sua comparação com o descritor de Zelnik [11] e o de Laptev [12].

A Seção 5.1 apresenta os primeiros estudos feitos com o descritor global proposto apresentado na Seção 4.1. Estes estudos foram realizados para descobrir se o descritor é viável para ser utilizado em um classificador SVM para o problema de reconhecimento de ações da base de vídeos KTH [10]. Como o resultado foi positivo, se tornou possível avaliar este descritor e suas variações.

Desta forma, a classificação da base foi feita utilizando um classificador SVM com validação cruzada. Esta avaliação é apresentada na Seção 5.2. Como não é objetivo deste trabalho, não se aprofundará no estudo do classificador SVM. Porém este assunto se mostra importante em futuros estudos sobre a influência das funções núcleo neste classificador.

O método de classificação faz parte do 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*) [33].

A Seção 5.3 apresenta a comparação entre os descritores propostos e os descritores de Zelnik [11] e de Laptev [12].

Todos os vídeos mostrados nos resultados pertencem à base de vídeo KTH [10]. Esta base de vídeo é composta por seis tipos de ações humanas:

- *Walking*: movimento no qual a pessoa anda para a direita ou para a esquerda;
- *Jogging*: movimento no qual a pessoa faz *jogging* para a direita ou para a esquerda;
- *Running*: movimento no qual a pessoa corre para a direita ou para a esquerda;
- *Boxing*: movimento no qual a pessoa dá socos no ar continuamente;
- *Hand waving*: movimento no qual a pessoa acena com ambas as mãos continuamente;

- *Hand clapping*: movimento no qual a pessoa bate palmas continuamente.

Cada uma dessas ações é realizada diversas vezes por 25 pessoas diferentes e em quatro cenários: ambiente externo (s1), ambiente externo com variação de escala (s2), ambiente externo com variação de velocidade (s3) e ambiente interno (s4) (Figura 1.1). No total, a base possui 2391 vídeos.

Para facilitar a apresentação dos resultados, nas Tabelas os movimentos são referenciados pelas seguintes abreviações: *walking* passa a ser **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**.

5.1 Estudo de viabilidade de uso

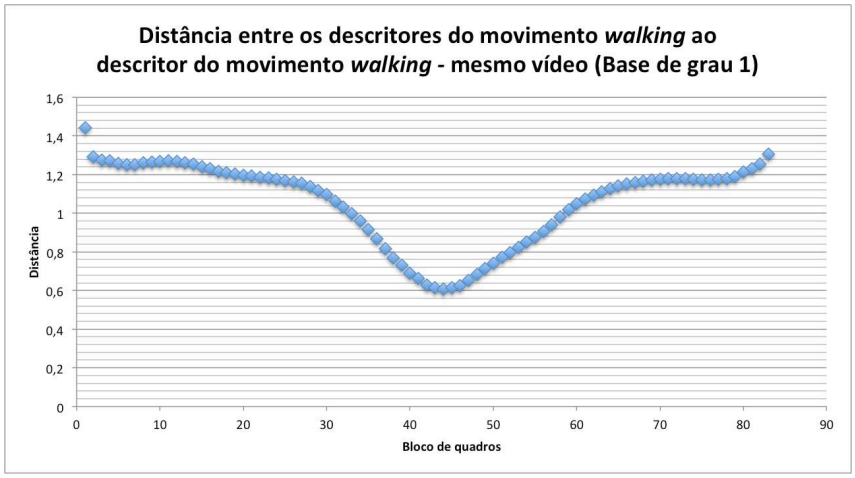
Estes estudos foram feitos para descobrir se o descritor proposto é viável para ser utilizado em um classificador SVM para o problema de reconhecimento de ações da base de vídeos KTH [10].

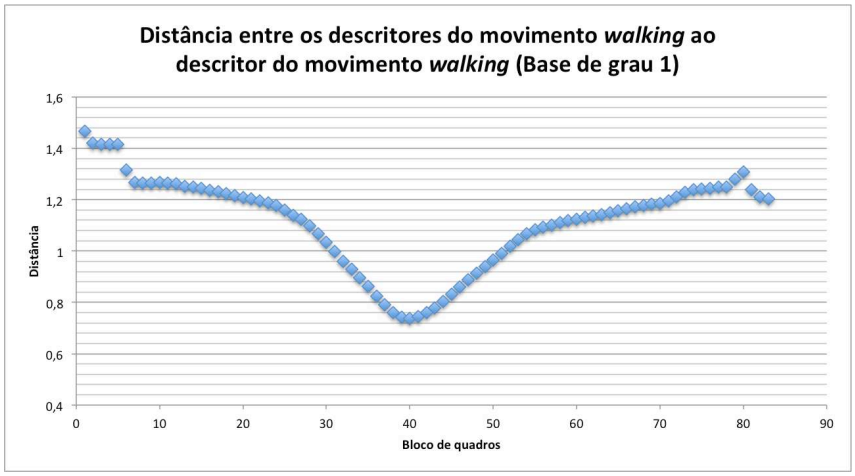
Em um primeiro momento foi criado um descritor $\bar{D}$ (apresentado na Seção 4.1) utilizando todos os quadros de um vídeo do movimento *walking*. Este descritor foi comparado a descritores formados por blocos de quadros de três tipos de vídeos: seu próprio vídeo, um vídeo diferente do movimento *walking* e um vídeo do movimento *boxing*. Os descritores são formados a cada bloco de 16 quadros do vídeo e são comparados calculando a distância l^2 (Eq. 4.5) entre cada um destes blocos e o descritor $\bar{D}$.

A Figura 5.1 apresenta o comportamento encontrado das distâncias calculadas quando os descritores são formados pelos coeficientes dos polinômios de uma base de grau 1.

É possível perceber que para o mesmo tipo de ação, o comportamento das distâncias entre os descritores são similares. Já entre vídeos de ações diferentes, o comportamento se mostra diferente. Este mesmo resultado aparece quando tem-se uma base de grau maior. A Figura 5.2 apresenta o comportamento das distâncias para uma base de grau 9.

É interessante destacar que existe uma pequena diferença entre o comportamento das distâncias quando aumenta-se o grau da base. De fato, quanto maior o grau da base, mais precisa é a modelagem do movimento e assim é possível detectar exatamente aonde ocorre o movimento de passo na cena. Assim, na Figura 5.2, observa-se, entre os quadros 30 e 60, que as distâncias dos vídeos do movimento *walking* oscilam, o que representa





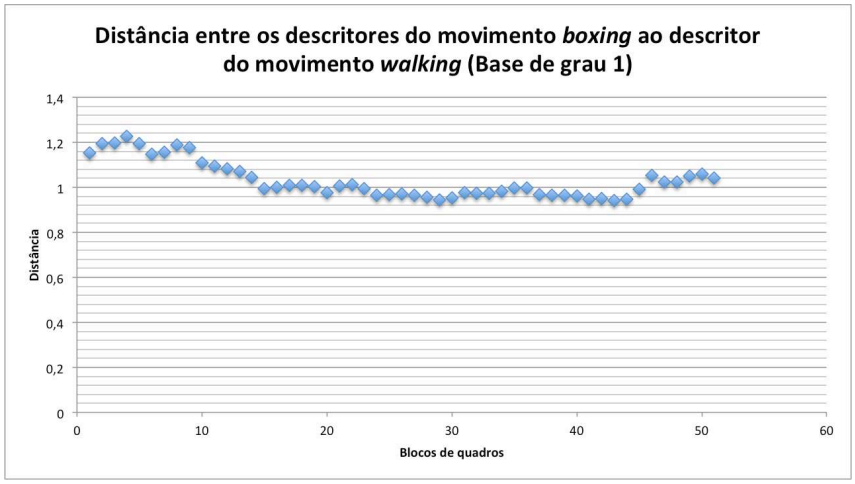


Figura 5.1: Descritor $\bar{D}$ de base de grau 1 comparado parte por parte com seu próprio vídeo, outro vídeo do movimento *walking* e um vídeo do movimento *boxing*

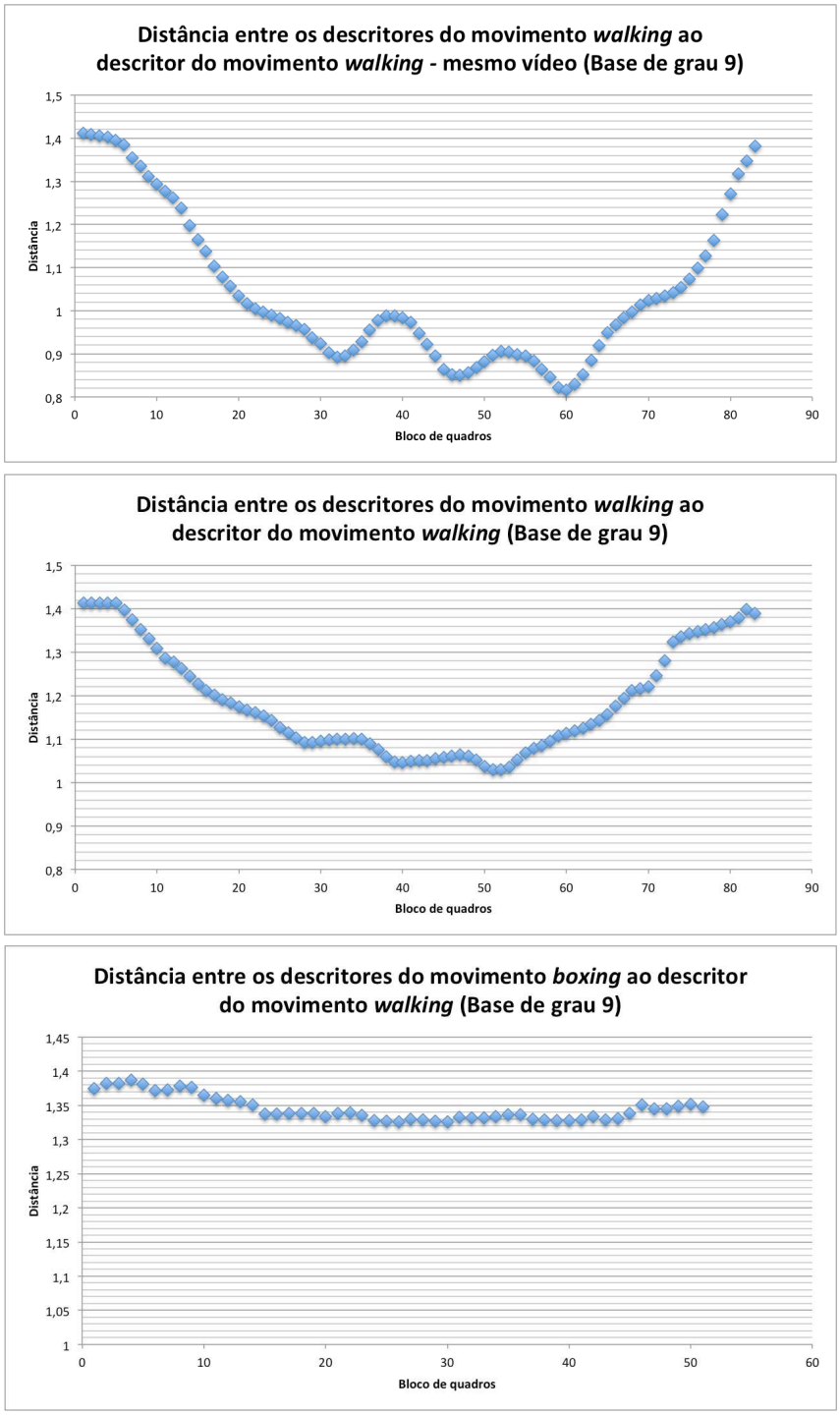


Figura 5.2: Descritor $\bar{D}$ de base de grau 9 comparado parte por parte com seu próprio vídeo, outro vídeo do movimento *walking* e um vídeo do movimento *boxing*

onde está acontecendo o movimento do tipo passo.

A partir deste resultado é possível perceber que o descritor criado é capaz de descrever movimentos diferentes. Porém, ainda não é possível concluir se o descritor é viável para ser utilizado em um classificador SVM. Para tal, foi montado um conjunto de teste com 192 vídeos da base KTH. A partir destes vídeos foram criados descritores para cada um deles e foi calculada a distância média l_2 (Eq. 4.5) e a variância entre os conjuntos da base. Para cada tipo de ação tem-se o mesmo número de vídeos.

A Tabela 5.1 mostra as distâncias médias e a Tabela 5.2 mostra as variâncias encontradas entre os conjuntos de vídeos utilizando uma base de grau 1.

	Walk	Box	HWav	HClap	Jog	Run
Walk	0,509	1,139	1,228	1,032	0,499	0,523
Box		0,696	0,737	0,832	1,121	1,106
HWav			0,524	0,835	1,216	1,195
HClap				0,733	1,011	0,982
Jog					0,476	0,510
Run						0,503

Tabela 5.1: Distâncias médias entre os conjuntos utilizando uma base de grau 1.

	Walk	Box	HWav	HClap	Jog	Run
Walk	0,062	0,030	0,008	0,029	0,053	0,046
Box		0,079	0,050	0,050	0,031	0,029
HWav			0,058	0,051	0,009	0,010
HClap				0,068	0,029	0,027
Jog					0,056	0,048
Run						0,052

Tabela 5.2: Variâncias entre os conjuntos utilizando uma base de grau 1.

Estas distâncias e variâncias mostram que o descritor consegue separar os movimentos *boxing*, *hand waving* e *hand clapping*, porém, os movimentos *walking*, *jogging* e *running* são considerados muito próximos. De fato, estes movimentos geram fluxos ópticos muito similares e uma base de grau 1 não é suficiente para descrevê-los. Uma forma de se resolver este problema é aumentar o grau da base, assim, os mesmos testes foram realizados para uma base de grau 9 (Tabelas 5.3 e 5.4).

As distâncias e variâncias das Tabelas 5.3 e 5.4 mostram que os resultados melhoraram e uma distância maior entre os conjuntos dos movimentos *walking*, *jogging* e *running* foi encontrada.

	Walk	Box	HWav	HClap	Jog	Run
Walk	1,067	1,307	1,313	1,283	1,165	1,225
Box		0,955	1,223	1,225	1,311	1,328
HWav			1,008	1,193	1,309	1,320
HClap				1,063	1,292	1,307
Jog					1,159	1,219
Run						1,198

Tabela 5.3: Distâncias médias entre os conjuntos utilizando uma base de grau 9

	Walk	Box	HWav	HClap	Jog	Run
Walk	0,067	0,008	0,005	0,011	0,024	0,016
Box		0,083	0,011	0,023	0,006	0,005
HWav			0,066	0,020	0,004	0,003
HClap				0,079	0,006	0,005
Jog					0,059	0,017
Run						0,059

Tabela 5.4: Variâncias entre os conjuntos utilizando uma base de grau 9

Desta forma, os resultados mostram que as distâncias médias entre conjuntos são separáveis, logo é possível utilizar o descritor tensorial proposto em um classificador SVM.

5.2 Aplicação na classificação de vídeos

Com os resultados obtidos pelos estudos iniciais, foi possível concluir que o descritor proposto é viável para a aplicação na classificação de vídeos utilizando um classificador SVM. Nesta seção serão apresentados os resultados da classificação da base KTH para cada uma das variações do descritor proposto apresentadas no Capítulo 4.

5.2.1 Descritor tensorial baseado em polinômios de Legendre que representam o fluxo óptico

O descritor apresentado na Seção 4.1 foi avaliado classificando a base KTH com um classificador SVM de função núcleo gaussiano L^2. A Figura 5.3 mostra a taxa de reconhecimento (ou precisão) encontrada para as bases de grau 1 à 9. Os valores exatos dessas precisões podem ser vistos na Tabela 5.5.

É interessante destacar que quanto maior o grau da base, melhor é a aproximação do fluxo óptico, porém se torna mais difícil a classificação, como vemos na Tabela 5.5 e na

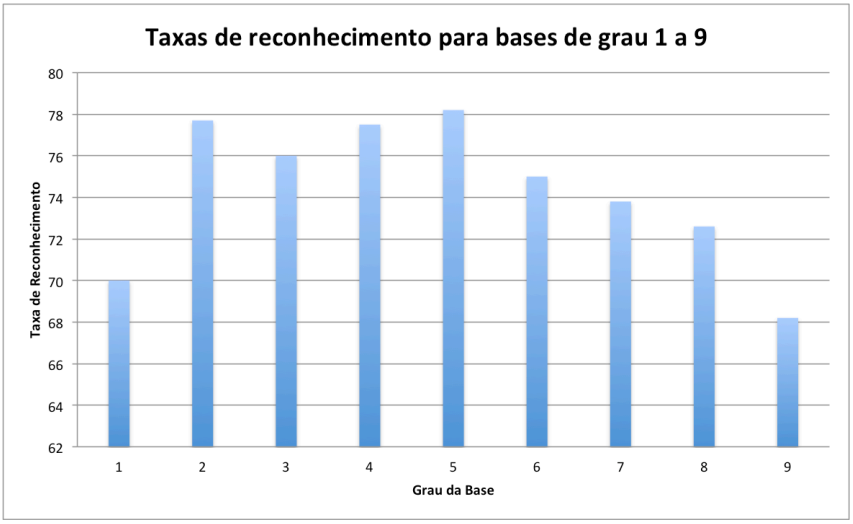


Figura 5.3: Taxas de Reconhecimento

Grau da Base	Taxas de reconhecimento
1	70%
2	77,7%
3	76%
4	77,5%
5	**78,2%**
6	75%
7	73,8%
8	72,6%
9	68,2%

Tabela 5.5: Taxas de reconhecimento

Figura 5.3.

O melhor resultado obtido foi para uma base de grau 5 ($78,2\%$). Para tentar aumentar esta precisão, os parâmetros do classificador foram trocados. Assim, com a mudança da função núcleo gaussiano para uma função núcleo triangular, a precisão encontrada foi de $79,96\%$. A matriz de confusão para este resultado é dada pela Tabela 5.6. Este tipo de matriz mostra a porcentagem que cada conjunto foi classificado em cada um dos tipos de conjunto. A direção de leitura desta matriz é a vertical.

A matriz de confusão (Tabela 5.6) mostra que a maior confusão de classificação é entre os conjuntos *walking*, *jogging* e *running*, sendo que o pior entre eles é o *jogging* com apenas $59,02\%$ dos vídeos desse movimento sendo classificados corretamente, $19,44\%$ classificados como *running* e $21,52\%$ classificados como *walking*.

	Box	HWav	HClap	Jog	Run	Walk
Box	**93,00%**	*5,55%*	*16,66%*	0,0%	0,0%	0,0%
HWav	*0,69%*	**82,63%**	*2,08%*	0,0%	0,0%	0,0%
HClap	*1,39%*	*11,80%*	**81,25%**	0,0%	0,0%	0,0%
Jog	*0,69%*	0,0%	0,0%	**59,02%**	*17,36%*	*9,72%*
Run	*4,19%*	0,0%	0,0%	*19,44%*	**77,77%**	*4,16%*
Walk	0,0%	0,0%	0,0%	*21,52%*	*4,86%*	**86,11%**

Tabela 5.6: Matriz de confusão para uma base de grau 5

5.2.2 Inserindo coerência temporal ao descritor

O descritor apresentado na Seção 4.2 foi avaliado classificando a base KTH com um classificador SVM de função núcleo triangular. A Figura 5.4 mostra a taxa de reconhecimento encontrada para as bases de grau 1 à 8. Os valores exatos dessas precisões podem ser vistos na Tabela 5.7.

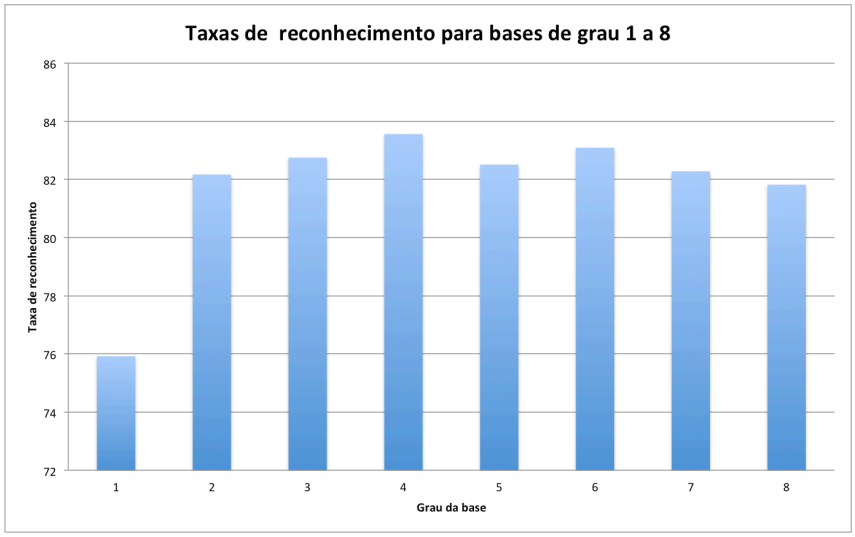


Figura 5.4: Taxas de Reconhecimento

Assim como os resultados apresentados na seção anterior, à partir de um certo grau da base, o descritor se torna cada vez mais específico e a precisão começa a decrescer. A melhor precisão encontrada por este descritor foi de 83,56% para uma base de grau 4. A matriz de confusão para este resultado é dada pela Tabela 5.8.

A matriz de confusão (Tabela 5.8) mostra que os movimentos com as mãos possuem menos problemas de classificação do que os movimentos do tipo passo. No caso, o pior problema de classificação está no movimento *jogging* com $64,58\%$ vídeos sendo classificados corretamente, $15,97\%$ sendo classificados como *running* e $19,44\%$ como *walking*.

Grau da Base	Taxas de reconhecimento
1	75,92%
2	82,17%
3	82,75%
4	**83,56%**
5	82,51%
6	83,09%
7	82,28%
8	81,82%

Tabela 5.7: Taxas de reconhecimento

	Box	HWav	HClap	Jog	Run	Walk
Box	**97,90%**	*12,50%*	*1,38%*	0,0%	0,0%	0,0%
HWav	0,0%	**87,50%**	*13,38%*	0,0%	0,0%	0,0%
HClap	0,0%	0,0%	**84,72%**	0,0%	0,0%	0,0%
Jog	0,0%	0,0%	0,0%	**64,58%**	*15,27%*	*11,11%*
Run	*1,39%*	0,0%	0,0%	*15,97%*	**79,86%**	*2,08%*
Walk	*0,69%*	0,0%	0,0%	*19,44%*	*4,86%*	**86,80%**

Tabela 5.8: Matriz de confusão para uma base de grau 4

Por ter apresentado uma melhor taxa de reconhecimento do que o descritor apresentado na Seção 4.1, os próximos testes utilizam o descritor com derivada.

5.2.3 Agrupamento de tensores no tempo

O descritor apresentado na Seção 4.3 foi avaliado classificando a base KTH com um classificador SVM de função núcleo triangular. Foram utilizadas as duas formas de divisão no tempo apresentadas na Figura 4.3. A Figura 5.5 mostra a taxa de reconhecimento encontrada para as bases de grau 1 a 7. Os valores exatos dessas precisões podem ser vistos na Tabela 5.9. O conjunto de descritores, com e sem redundância, é formado por tensores com informação de derivada.

Pode-se perceber que da mesma forma que nos testes anteriores, a partir de um certo grau da base a taxa de reconhecimento começa a decrescer. A melhor precisão encontrada por este descritor foi de $81,82\%$ para uma base de grau 4 com 3 divisões no tempo. A matriz de confusão para este resultado é dada pela Tabela 5.10.

O principal problema de classificação está no movimento *jogging* com $63,19\%$ dos vídeos sendo classificados corretamente, $18,75\%$ como *running* e $18,05\%$ como *walking*.

A partir destes resultados é possível perceber que não houveram grandes diferenças

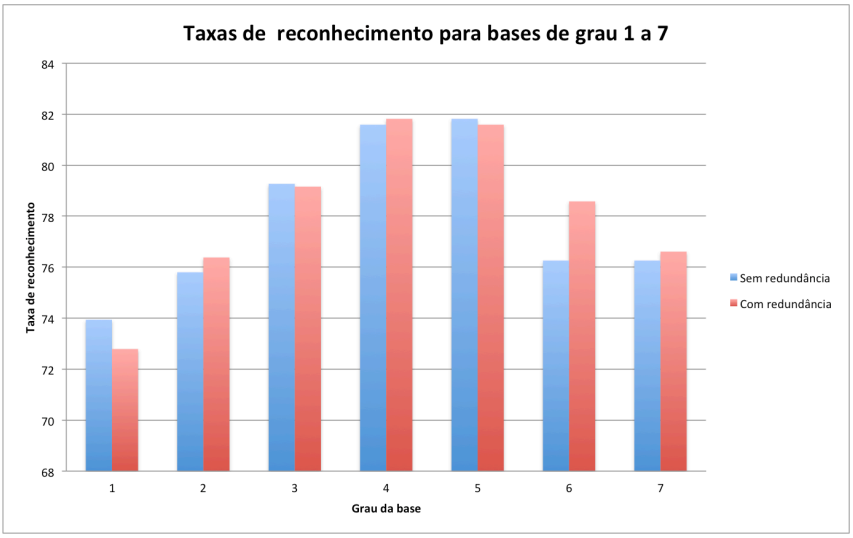


Figura 5.5: Taxas de Reconhecimento

	Sem redundância	Com redundância
Grau da Base	Taxas de reconhecimento	Taxas de reconhecimento
1	73,94%	72,79%
2	75,80%	76,38%
3	79,27%	79,16%
4	81,59%	**81,82%**
5	81,82%	81,59%
6	76,26%	78,58%
7	76,26%	76,61%

Tabela 5.9: Taxas de reconhecimento

entre as divisões com e sem redundância. Isso mostra que o reforço no intervalo de quadros $[\frac{n_f}{4}, \frac{3n_f}{4}]$ não é válido e não introduz muita informação ao descritor.

5.2.4 Agrupamento de tensores no espaço

O descritor apresentado na Seção 4.4 foi avaliado classificando a base KTH com um classificador SVM de função núcleo triangular. Foram utilizadas as três formas de divisão no espaço apresentadas na Figura 4.4. A Figura 5.6 mostra a taxa de reconhecimento encontrada para as bases de grau 1 à 7. Os valores exatos dessas precisões podem ser vistos na Tabela 5.11. O conjunto de descritores é formado por tensores com informação de derivada.

É possível perceber que o comportamento da taxa de reconhecimento é o mesmo encontrado nos testes anteriores, isto é, a partir de um certo grau da base, a taxa de

	Box	HWav	HClap	Jog	Run	Walk
Box	**97,10%**	*11,80%*	*1,38%*	0,0%	0,0%	0,0%
HWav	*0,69%*	**84,72%**	*15,97%*	0,0%	0,0%	0,0%
HClap	0,0%	0,0%	**82,63%**	0,0%	0,0%	0,0%
Jog	*1,39%*	0,0%	0,0%	**63,19%**	*13,19%*	*11,11%*
Run	*2,09%*	*3,47%*	0,0%	*18,75%*	**81,94%**	*5,55%*
Walk	*0,69%*	0,0%	0,0%	*18,05%*	*4,86%*	**83,33%**

Tabela 5.10: Matriz de confusão para uma base de grau 4

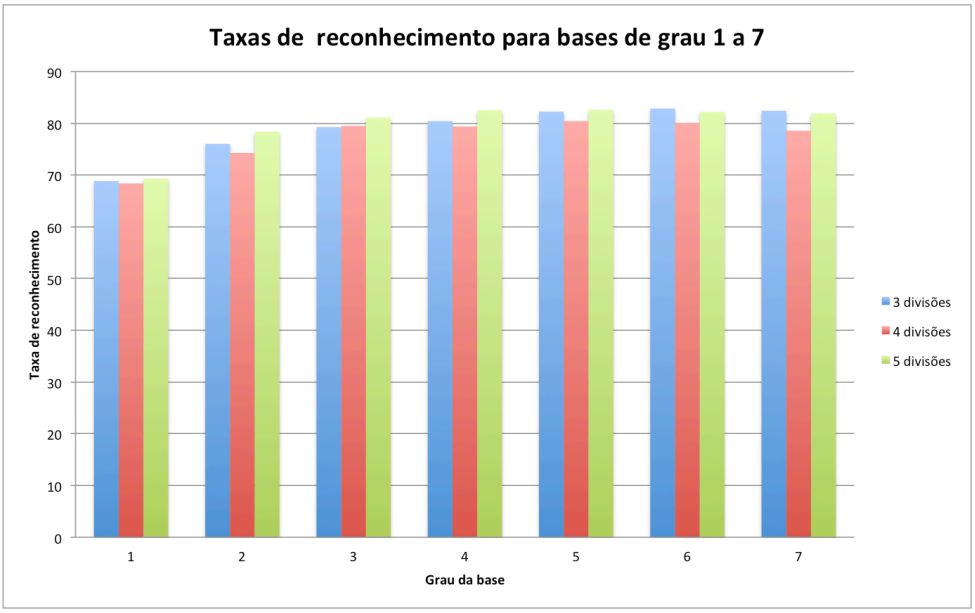


Figura 5.6: Taxas de Reconhecimento

reconhecimento começa a cair. Percebe-se também que o conjunto de descritores formado a partir da divisão no espaço com redundância (5 divisões no espaço) se mostrou mais eficiente do que o conjunto formado por 4 divisões no tempo, porém, a melhor precisão ficou abaixo da obtida com 3 divisões no tempo. Isso mostra que o reforço na região central dos quadros do vídeo é válido se comparado com 4 divisões no tempo, mas não supera as 3 divisões no tempo.

O melhor resultado obtido foi de $82,86\%$ com 3 divisões no tempo e base de grau 6. A matriz de confusão para este resultado é dada pela Tabela 5.12.

O principal problema de classificação está no movimento *jogging* com $63,19\%$ dos vídeos sendo classificados corretamente, $15,97\%$ como *running* e $20,93\%$ como *walking*.

	Divisões		
	3	4	5
Grau da Base	Taxas	Taxas	Taxas
1	68,84%	68,39%	69,32%
2	76,03%	74,28%	78,35%
3	79,27%	79,50%	81,13%
4	80,43%	79,39%	82,51%
5	82,28%	80,43%	82,63%
6	**82,86%**	80,08%	82,16%
7	82,44%	78,57%	81,93%

Tabela 5.11: Taxas de reconhecimento

	Box	HWav	HClap	Jog	Run	Walk
Box	**97,10%**	*11,80%*	*1,38%*	0,0%	0,0%	0,0%
HWav	0,0%	**87,50%**	*13,88%*	0,0%	0,0%	*0,69%*
HClap	*2,09%*	0,0%	**84,72%**	0,0%	0,0%	0,0%
Jog	*1,38%*	0,0%	0,0%	**63,19%**	*18,05%*	*7,63%*
Run	*1,39%*	*0,69%*	0,0%	*15,97%*	**77,77%**	*2,77%*
Walk	0,0%	0,0%	0,0%	*20,93%*	*4,16%*	**88,88%**

Tabela 5.12: Matriz de confusão para uma base de grau 6

5.2.5 Descritor obtido em janela deslizante

O descritor apresentado na Seção 4.5 foi avaliado classificando a base KTH com um classificador SVM de função núcleo triangular. A Figura 5.7 mostra a taxa de reconhecimento encontrada para as bases de grau 1 a 8 com uma janela deslizante de dimensão 60×100. Os valores exatos dessas precisões podem ser vistos na Tabela 5.13. O descritor é formado com informação de derivada.

Grau da Base	Taxa de Reconhecimento
1	76,95%
2	84,94%
3	84,95%
4	84,25%
5	**85,52%**
6	83,67%
7	83,67%
8	82,74%

Tabela 5.13: Taxas de reconhecimento

Da mesma forma como encontrado nos testes anteriores, a partir de um certo grau

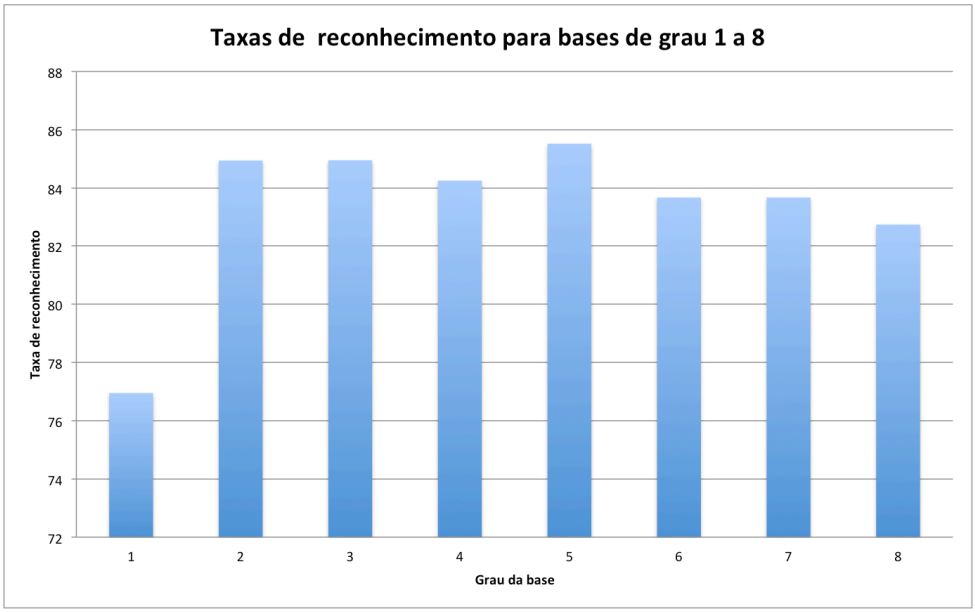


Figura 5.7: Taxas de Reconhecimento

da base a taxa de reconhecimento começa a diminuir. Para tentar aumentar a taxa de reconhecimento foram testados outros tamanhos de janela. As melhores precisões e o tamanho da janela usada são apresentados na Tabela 5.14.

Tamanho da Janela	Grau da Base	Taxa de Reconhecimento
30×60	3	79,62%
40×90	3	83,79%
50×90	3	85,29%
40×100	3	84,25%
50×100	**3**	**86,80%**

Tabela 5.14: Taxas de reconhecimento para diferentes tamanhos de janela

O melhor resultado obtido com o descritor proposto foi de $86,80\%$ para o descritor obtido com uma janela deslizante de dimensão 50×100 e base de grau 3. A matriz de confusão para esse resultado é dada pela Tabela 5.15.

A maior confusão de classificação é com o movimento *running*. Para este movimento, $77,77\%$ dos vídeos foram classificados corretamente, $0,69\%$ foram classificados como *boxing*, $14,58\%$ como *jogging* e $6,94\%$ como *walking*.

	Box	HClap	HWav	Jog	Run	Walk
Box	**95,80%**	*6,94%*	0,0%	*1,38%*	*0,69%*	0,0%
HClap	0,0%	**90,27%**	*16,66%*	0,0%	0,0%	0,0%
HWav	0,0%	*2,08%*	**83,33%**	0,0%	0,0%	0,0%
Jog	*0,69%*	0,0%	0,0%	**79,86%**	*14,58%*	*2,08%*
Run	0,0%	0,0%	0,0%	*3,47%*	**77,77%**	*4,16%*
Walk	*3,49%*	*0,69%*	0,0%	*15,27%*	*6,94%*	**93,75%**

Tabela 5.15: Matriz de confusão encontrada pelo descritor proposto obtido em janela deslizante

5.2.6 Considerações

A Figura 5.8 mostra as melhores precisões encontradas por cada um dos descritores criados neste trabalho. Na Figura, o termo *original* representa o descritor apresentado na Seção 4.1, *com derivada* o descritor apresentado na Seção 4.2, *divisões no tempo* o descritor apresentado na Seção 4.3, *divisões no espaço* o descritor apresentado na Seção 4.4 e *em janela* o descritor apresentado na Seção 4.5. Os valores exatos dessas precisões podem ser vistos na Tabela 5.16.

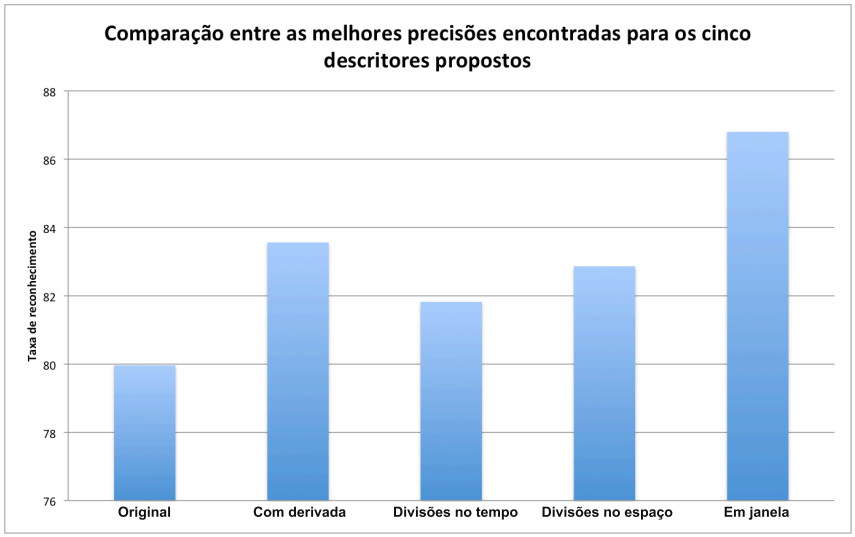


Figura 5.8: Comparação entre os cinco descritores propostos.

Observando as precisões dos dois primeiros descritores, é possível afirmar que a velocidade (derivada de primeira ordem dos coeficientes) adiciona muita informação ao descritor, uma vez que a taxa de reconhecimento encontrada utilizando o descritor com derivada se mostrou bastante superior à encontrada pelo descritor original. Desta forma, todos os outros testes foram feitos com tensores com informação de derivada.

Tipo do descritor	Taxa de Reconhecimento
Original	79,96%
Com derivada	83,56%
Divisões no tempo	81,82 %
Divisões no espaço	82,86 %
Em janela	86,80 %

Tabela 5.16: Comparação entre os cinco descritores propostos.

O uso de pacote de características com divisões no espaço apresentou uma precisão melhor do que a encontrada pelo descritor com divisões no tempo.

É importante destacar que a taxa de reconhecimento encontrada pelos descritores que usam pacote de características é inferior à do descritor com derivada. Isso mostra que é necessário estudar outras formas de criação de descritores usando essa abordagem.

Obtendo o descritor a partir de uma janela deslizante, em vez de se aproximar todo o fluxo óptico presente no quadro, aproxima-se apenas a região de interesse que contém o movimento mais representativo do quadro. Com efeito, observa-se na Figura 5.8 que a precisão encontrada pelo descritor obtido em janela deslizante é bastante superior às outras precisões encontradas. Isso pode ser explicado porque ao se escolher uma região de interesse com os movimentos mais representativos, diminui-se o ruído presente no quadro, melhorando a aproximação do fluxo óptico e, por consequência, o descritor tensorial passa a representar melhor o movimento da cena.

De maneira geral, a partir de um certo grau da base, a precisão tende a diminuir. De fato, quanto maior o grau da base, maior é o tamanho do descritor, o que pode atrapalhar a classificação. A Figura 5.9 mostra o comportamento de crescimento exponencial do tamanho do descritor à medida que se aumenta o grau da base. Em função disto, quanto mais se aumenta o grau da base, mais o descritor se torna específico e muito diferente de todos os outros descritores.

Em todos os casos, as ações com maiores problemas de classificação foram *walking*, *jogging* e *running*. Isto se explica por estes movimentos terem fluxo óptico semelhante nas imagens projetadas, o que complica a classificação. É importante destacar que este tipo de confusão é encontrado em diversos trabalhos presentes na literatura.

Dentre estes movimentos, o que apresentou uma menor taxa de reconhecimento e uma maior confusão na maioria dos descritores propostos foi o movimento *jogging*. De fato, este tipo de ação é um movimento de velocidade intermediária entre *walking* e *running*, assim,

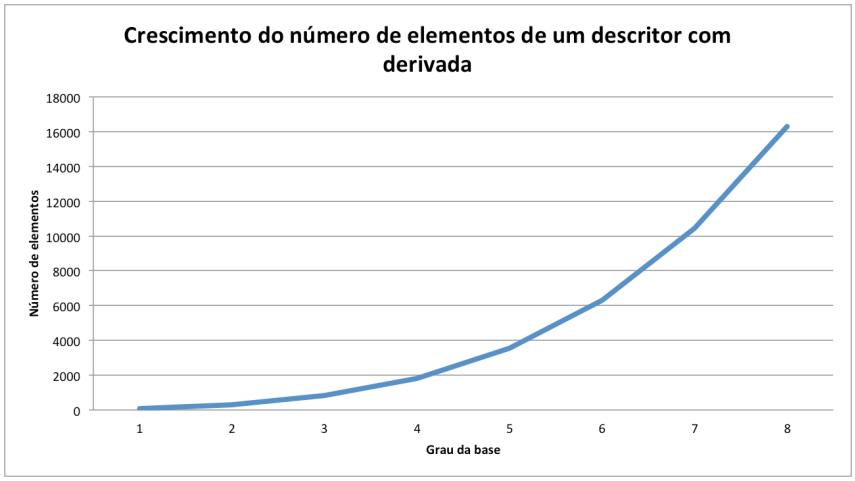


Figura 5.9: Comportamento do número de elementos do descritor com derivada à medida que se aumenta o grau da base.

se uma pessoa anda mais rápido, seu movimento *jogging* pode ser considerado rápido demais e ser classificado como *running*. Da mesma forma, se uma pessoa anda muito devagar, seu movimento *jogging* pode ser considerado devagar demais e ser classificado como *walking*.

5.3 Comparação com descritores globais da literatura

A implementação dos descritores de Zelnik *et al* [11] e de Laptev *et al* [12] se mostrou necessária para que eles sejam avaliados nas mesmas condições dos descritores propostos.

Em seu trabalho, Laptev *et al* [12] apresenta apenas a precisão encontrada para a classificação da base KTH utilizando um classificador SVM para o seu descritor estendido (3 escalas temporais e 3 escalas espaciais). Esta precisão pode ser vista em destaque na Figura 5.10 e é aproximadamente 71%.

A versão do descritor de Zelnik *et al* [11] implementada neste trabalho alcançou uma taxa de reconhecimento de 63,85%. Já a versão implementada do descritor de Laptev *et al* [12] alcançou 70,66%, o que está próximo da precisão apresentada em seu artigo.

Como pode se perceber, o descritor proposto nesta dissertação e suas variações alcançam todos precisões superiores às encontradas pelos descritores globais presentes na literatura (Figura 5.11).

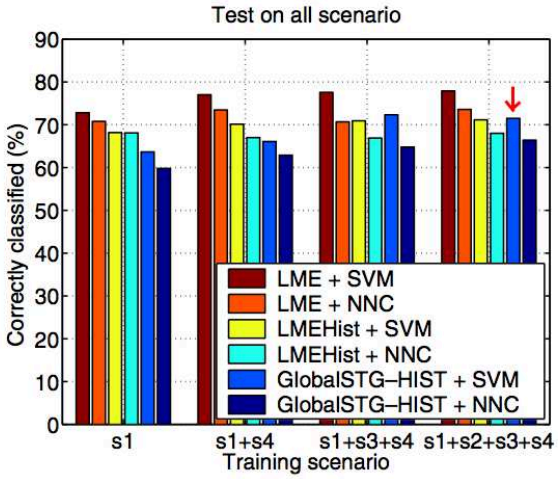


Figura 5.10: Precisões encontradas por Laptev *et al* [12]

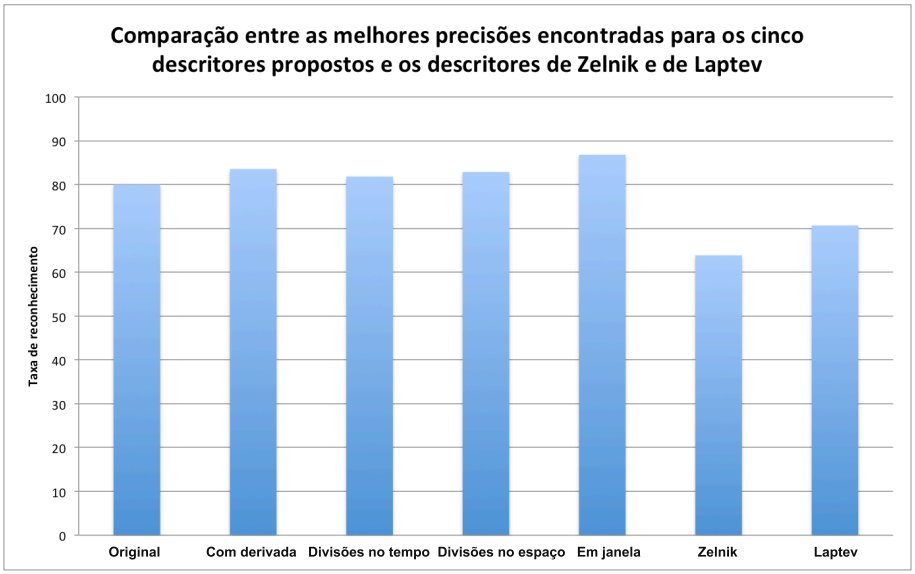


Figura 5.11: Comparação entre os cinco descritores propostos e os descritores de Zelnik *et al* [11] e de Laptev *et al* [12]

Estes resultados mostram que um descritor baseado no tensor de orientação consegue capturar e descrever melhor o movimento presente na cena quando comparado a um descritor baseado em histograma de gradientes. De fato, o histograma de gradientes só leva em consideração a quantidade de direções presente na cena, enquanto o descritor proposto se baseia no fluxo óptico e sua variação no tempo.

A matriz de confusão para o resultado obtido pelo descritor de Zelnik *et al* [11] é dada pela Tabela 5.17 e pelo descritor de Laptev *et al* [12] é dada pela Tabela 5.18.

	Box	HClap	HWav	Jog	Run	Walk
Box	**68,53%**	*44,44%*	*27,77%*	*1,38%*	*0,69%*	*3,47%*
HClap	*19,58%*	**49,30%**	*14,58%*	0,0%	*3,47%*	0,0%
HWav	*9,79%*	*3,47%*	**54,86%**	*2,08%*	0,0%	*2,08%*
Jog	0,0%	*1,38%*	*0,69%*	**54,16%**	*11,80%*	*22,22%*
Run	0,0%	*0,69%*	0,0%	*31,94%*	**81,94%**	*0,69%*
Walk	*2,09%*	*0,69%*	*2,08%*	*10,41%*	*2,08%*	**71,52%**

Tabela 5.17: Matriz de confusão encontrada pelo descritor de Zelnik *et al* [11]

	Box	HClap	HWav	Jog	Run	Walk
Box	**58,04%**	*11,11%*	*14,58%*	*1,38%*	0,0%	0,0%
HClap	*20,27%*	**72,91%**	*16,66%*	0,0%	0,0%	0,0%
HWav	*17,48%*	*15,27%*	**65,97%**	*2,08%*	*1,38%*	*2,77%*
Jog	0,0%	0,0%	0,0%	**52,77%**	*10,41%*	*10,41%*
Run	*2,09%*	*0,69%*	0,0%	*29,86%*	**87,5%**	0,0%
Walk	*2,09%*	*0,69%*	*2,77%*	*13,88%*	*0,69%*	**86,80%**

Tabela 5.18: Matriz de confusão encontrada pelo descritor de Laptev *et al* [12]

A matriz de confusão encontrada pelo descritor de Zelnik *et al* [11] mostra que o maior problema de classificação está no movimento *handclapping*, no qual $49,30\%$ dos vídeos deste tipo de movimento foram classificados corretamente, $44,44\%$ foram classificados como sendo do movimento *boxing*, $3,47\%$ como sendo do movimento *hand waving*, $1,38\%$ como sendo do movimento *jogging*, $0,69\%$ como sendo do movimento *running* e $0,69\%$ como sendo do movimento *walking*.

A inserção das escalas espaciais de Laptev *et al* [12] melhoraram a precisão encontrada e diminuíram consideravelmente a confusão de classificação do movimento *handclapping*. Neste caso, a pior classificação está no movimento *jogging*, com $52,77\%$ dos vídeos sendo classificados corretamente, $29,86\%$ como sendo do movimento *running*, $13,88\%$ como

sendo do movimento *walking*, $2,08\%$ como sendo do movimento *hand waving* e $1,38\%$ como sendo do movimento *boxing*. De maneira geral, mesmo como o movimento *jogging* tendo a pior classificação, os movimentos de braço apresentam maiores confusões de classificação.

É possível concluir que ambos os descritores baseados em histogramas de gradiente possuem maior problema de classificação entre movimentos de braço.

O melhor resultado obtido com o descritor proposto foi de $86,80\%$ para o descritor obtido com uma janela deslizante de dimensão 50×100 e base de grau 3. A matriz de confusão para esse resultado é dada pela Tabela 5.15, apresentada na Seção 5.2.5.

É possível concluir que o descritor tensorial tem menos problemas de classificação e se confunde mais entre os movimentos do tipo passo, no caso, a maior confusão é com o movimento *running*. Para este movimento, $77,77\%$ dos vídeos foi classificado corretamente, $0,69\%$ foi classificado como *boxing*, $14,58\%$ como *jogging* e $6,94\%$ como *walking*.

6 CONCLUSÕES E PERSPECTIVAS

Neste trabalho foi apresentado um novo descritor global de movimento. Este descritor é baseado no tensor de orientação formado à partir dos coeficientes dos polinômios de Legendre calculados para cada quadro de um vídeo. Os coeficientes são encontrados através da projeção do fluxo óptico nos polinômios de Legendre, obtendo-se uma representação polinomial do movimento.

O uso de polinômios de Legendre que representam o fluxo óptico, sua codificação na forma de tensores e a proposta de acumulá-los para a criação do descritor é a principal contribuição deste trabalho.

Com o objetivo de se melhorar o desempenho do descritor em aplicações de classificação, foram propostas quatro variações do descritor: adicionando a informação de velocidade (derivada de primeira ordem em relação ao tempo), usando pacote de características dividindo o vídeo no tempo e dividindo o vídeo no espaço, e obtendo o descritor em uma janela deslizante.

Pelo fato de o descritor ser uma acumulação de tensores, as principais limitações do descritor são a descontinuidade do movimento e um número elevado de quadros do vídeo. Havendo um desses casos, o descritor tensorial tende a se tornar isotrópico, não representando corretamente a informação de direção do movimento.

O descritor tensorial criado foi avaliado classificando-se a base de vídeos KTH [10] com um classificador SVM (máquina de vetor de suporte). A melhor taxa de reconhecimento encontrada foi de $86,80\%$ com o descritor obtido em janela deslizante utilizando uma base de grau 3.

Os maiores problemas de classificação, para todas as variações do descritor proposto, são entre os vídeos dos movimentos *walking*, *jogging* e *running*. Isto se explica por estes movimentos terem fluxo óptico semelhante nas imagens projetadas. Este tipo de problema é recorrente em diversos trabalhos da literatura.

O tipo de ação *jogging* é o movimento que apresenta mais problemas de classificação, já que este tipo de ação é um movimento de velocidade intermediária entre *walking* e

running, assim, se uma pessoa anda mais rápido, seu movimento *jogging* pode ser considerado rápido demais e ser classificado como *running*. Da mesma forma, se uma pessoa anda muito devagar, seu movimento *jogging* pode ser considerado devagar demais e ser classificado como *walking*.

Comparando o descritor proposto com dois descritores globais, baseados em histograma de gradientes (os descritores de Zelnik *et al* [11] e de Laptev *et al* [12]), foi possível concluir que a precisão da abordagem deste trabalho superou às encontradas por estes descritores globais. De fato, o histograma de gradientes só leva em consideração a quantidade de direções presente na cena, enquanto o descritor proposto se baseia no fluxo óptico e em sua variação no tempo, carregando assim mais informação.

Quando comparado a descritores locais, presentes na literatura, o descritor proposto ainda não supera a melhor taxa de reconhecimento encontrada (Tabela 6.1).

Abordagem	Taxa de Reconhecimento
Kim *et al* [27]	95,33%
Ikizler *et al* [24]	94%
Baysal *et al* [26]	91,5%
Ballan *et al* [25]	91,2%
Descritor tensorial proposto	**86,80%**
Shuldt *et al* [10]	71,7%

Tabela 6.1: Comparação entre descritores locais da literatura e o descritor tensorial proposto.

Dessa forma, se faz necessário estudos sobre como aumentar a precisão do descritor. Uma idéia para tal, mantendo o descritor global, é estudar as características espectrais do tensor, uma vez que as informações contidas nos autovalores e autovetores do tensor podem revelar características importantes do movimento. Outra idéia é estudar outras formas de se criar um descritor utilizando pacote de características. Caso ainda assim não ultrapasse as precisões encontradas pelos descritores locais, é possível combinar o descritor proposto com outros descritores, como por exemplo descritores locais, perdendo assim sua propriedade global.

No entanto, enquanto descritor puramente global, o descritor tensorial proposto se mostrou muito eficiente e promissor.

Pretende-se, ainda, avaliar o descritor em bases de vídeos mais complexas, como por exemplo, bases com vídeos que tenham mais de um movimento na cena.

REFERÊNCIAS

[1] AUGEREAU, B., TREMBLAIS, B., FERNANDEZ-MALOIGNE, C., "Vectorial Computation of the Optical Flow in Color Image Sequences." In: *Thirteenth Color Imaging Conference*, pp. 130–134, November 2005.

[2] LAUZE, F., KORNPROBST, P., LENGLET, C., DERICHE, R., NIELSEN, M., "Sur Quelques Méthodes de Calcul de Flot Optique à partir du Tenseur de Structure : Synthèse et Contribution". 2004.

[3] DRUON, M., TREMBLAIS, B., AUGEREAU, B., "Modélisation de champs de vecteurs par bases de polynômes : application à l'analyse de la posture d'utilisateurs devant un écran d'ordinateur, via une webcam." In: *COmpression et REprésentation des Signaux Audiovisuels (CORESA)*, pp. 290–295, Caen, France, Novembre 2006.

[4] LI, X., ZHAO, X., FU, Y., LIU, Y., "Bimodal gender recognition from face and fingerprint", *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, v. 0, pp. 2590–2597, 2010.

[5] MÜLLER, H., MICHOUX, N., BANDON, D., GEISSBUHLER, A., "A Review of Content-Based Image Retrieval Systems in Medical Applications – Clinical Benefits and Future Directions", *International Journal of Medical Informatics*, 2004.

[6] MOSLAH, O., VOINEA, D., NKOUTCHE-TEDJONG, R., SANCHEZ-GUINEA, A., KHIAL, Y., MOIGNE, V. L., COUVET, S., "A Model-Based Facade Reconstruction Approach using Shape Grammars", *ACM Transactions on Graphics*, 2010.

[7] REN, Y., GU, C., "Real-time hand gesture recognition based on vision". In: *Proceedings of the Entertainment for education, and 5th international conference on E-learning and games*, *Edutainment'10*, pp. 468–475, Springer-Verlag: Berlin, Heidelberg, 2010.

[8] GAO, X., YANG, Y., TAO, D., LI, X., "Discriminative optical flow tensor for video semantic analysis", *Comput. Vis. Image Underst.*, v. 113, pp. 372–383, March 2009.

[9] DRUON, M., *Modélisation du mouvement par polynômes orthogonaux : application à l'étude d'écoulements fluides*, Ph.D. Thesis, Université de Poitiers, 02 2009.

[10] SCHÜLDT, C., LAPTEV, I., CAPUTO, B., "Recognizing human actions: A local SVM approach". In: *In Proc. ICPR*, pp. 32–36, 2004.

[11] ZELNIK-MANOR, L., IRANI, M., "Event-based analysis of video". In: *In Proc. CVPR*, pp. 123–130, 2001.

[12] LAPTEV, I., CAPUTO, B., SCHULDT, C., LINDEBERG, T., "Local velocity-adapted motion events for spatio-temporal recognition", *Comput. Vis. Image Underst.*, v. 108, pp. 207–229, December 2007.

[13] LUCAS, B. D., KANADE, T., "An Iterative Image Registration Technique with an Application to Stereo Vision (IJCAI)". In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI 81)*, pp. 674–679, April 1981.

[14] HORN, B. K., SCHUNCK, B. G., *Determining Optical Flow*, Tech. rep., Cambridge, MA, USA, 1980.

[15] BARRON, J. L., FLEET, D. J., BEAUCHEMIN, S. S., "Performance of optical flow techniques", *INTERNATIONAL JOURNAL OF COMPUTER VISION*, v. 12, pp. 43–77, 1994.

[16] HAYKO, R., MICHAEL, D., HORST, B., "Bag of Optical Flow Volumes for Image Sequence Recognition". Proceedings of British Machine Vision Conference (BMVC), 2009, 2009.

[17] ZACH, C., POCK, T., BISCHOF, H., "A Duality Based Approach for Realtime TV-L1 Optical Flow". pp. 214–223, Proceedings of Symposium of the German Association for Pattern Recognition (DAGM), 2007.

[18] LOPES, A., OLIVEIRA, R., ALMEIDA, J., ARAÚJO, A. D. A., "Comparing alternatives for capturing dynamic information in bag-of-visual-features approaches

applied to human actions recognition". In: *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, IEEE Press, 2009.

[19] LOWE, D. G., "Object Recognition from Local Scale-Invariant Features", *Computer Vision, IEEE International Conference on*, v. 2, pp. 1150–1157 vol.2, Aug. 1999.

[20] BAY, H., ESS, A., TUYTELAARS, T., VAN GOOL, L., "Speeded-Up Robust Features (SURF)", *Comput. Vis. Image Underst.*, v. 110, pp. 346–359, June 2008.

[21] LAPTEV, I., LINDEBERG, T., "Space-time Interest Points". In: *IN ICCV*, pp. 432–439, 2003.

[22] LOPES, A. P. B., OLIVEIRA, R. S., DE ALMEIDA, J. M., DE A. ARAUJO, A., "Spatio-Temporal Frames in a Bag-of-Visual-Features Approach for Human Actions Recognition", *Computer Graphics and Image Processing, Brazilian Symposium on*, v. 0, pp. 315–321, 2009.

[23] BLANK, M., GORELICK, L., SHECHTMAN, E., IRANI, M., BASRI, R., "Actions as Space-Time Shapes". In: *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pp. 1395–1402, 2005.

[24] IKIZLER, N., CINBIS, R. G., DUYGULU, P., "Human action recognition with line and flow histograms". In: *In Proc. ICPR*, 2008.

[25] BALLAN, L., BERTINI, M., DEL BIMBO, A., SEIDENARI, L., SERRA, G., "Recognizing human actions by fusing spatio-temporal appearance and motion descriptors". In: *ICIP'09: Proceedings of the 16th IEEE international conference on Image processing*, pp. 3533–3536, IEEE Press: Piscataway, NJ, USA, 2009.

[26] BAYSAL, S., KURT, M. C., DUYGULU, P., "Recognizing Human Actions Using Key Poses", *Pattern Recognition, International Conference on*, v. 0, pp. 1727–1730, 2010.

[27] KYUN KIM, T., FAI WONG, S., CIPOLLA, R., "R.: Tensor Canonical Correlation Analysis for Action Classification". In: *In: CVPR 2007*, 2007.

[28] KRAUSZ, B., BAUCKHAGE, C., "Action Recognition in Videos Using Nonnegative Tensor Factorization", *Pattern Recognition, International Conference on*, v. 0, pp. 1763–1766, 2010.

[29] JIA, C., WANG, S., XU, X., ZHOU, C., ZHANG, L., "Tensor analysis and multi-scale features based multi-view human action recognition". In: *Proceedings of the 2nd International Conference on Computer Engineering and Technology*, *ICCET'10*, IEEE Press, 2010.

[30] 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*, *CIRA'09*, pp. 398–403, IEEE Press: Piscataway, NJ, USA, 2009.

[31] WESTIN, C.-F., *A Tensor Framework for Multidimensional Signal Processing*, Ph.D. Thesis, Linköping University, Sweden, S-581 83 Linköping, Sweden, 1994, Dissertation No 348, ISBN 91-7871-421-4.

[32] TORRES, R. D. S., AO, A. X. F., "Content-Based Image Retrieval: Theory and Applications", *Revista de Informática Teórica e Aplicada*, v. 13, pp. 161–185, 2006.

[33] FOURNIER, J., CORD, M., PHILIPP-FOLIGUET, S., "RETIN: A Content-Based Image Indexing and Retrieval System", *Pattern Analysis & Applications*, v. 4, n. 2, pp. 153–173, June 2001.