Pós-Graduação em Ciência da Computação

“Classificação de Esportes em Vídeos Amadores e Profissionais”

Por

Guilherme Ramalho Magalhães

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, 2014

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GUILHERME RAMALHO MAGALHÃES

“Classificação de Esportes em Vídeos Amadores e Profissionais"

ESTE TRABALHO FOI APRESENTADO À PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIA DA COMPUTAÇÃO.

ORIENTADOR: Prof. Tsang Ing Ren

RECIFE, 2014

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

M188c Magalhães, Guilherme Ramalho
Classificação de esportes em vídeos amadores e profissionais. / Guilherme Ramalho Magalhães. – Recife: O Autor, 2014.
62 f.: il., fig., tab.

Orientador: Tsang Ing Ren.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, 2014.
Inclui referências.

1. Inteligência artificial. 2. Aprendizagem de máquina. 3. Visão computacional. I. Ren, Tsang Ing (orientador). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2014-188

Dissertação de Mestrado apresentada por **Guilherme Ramalho Magalhães** à Pós Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título "**Classificação de Esportes em Vídeos Amadores e Profissionais**" orientada pelo **Prof. Tsang Ing Ren** e aprovada pela Banca Examinadora formada pelos professores:

_____
Prof. Paulo Jorge Leitão Adeodato
Centro de Informática/UFPE

_____
Prof. Leonardo Vidal Batista
Departamento de Informática/UFPB

_____
Prof. Tsang Ing Ren
Centro de Informática /UFPE

Visto e permitida a impressão.
Recife, 26 de agosto de 2014.

_____
Profa. Edna Natividade da Silva Barros
Coordenadora da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.

Dedico este trabalho aos meus familiares, amigos e professores que me ajudaram a chegar até aqui.

Agradecimentos

Uma série de agradecimentos faz-se necessária diante da conclusão desse trabalho. Agradeço primeiramente a Deus pelas oportunidades colocadas em minha vida; à minha família pela compreensão de minhas ausências e pela força nas horas mais difíceis; aos professores e especialmente ao meu orientador e co-orientador por em nenhum momento duvidar das minhas capacidades e mostrar que sempre podemos fazer melhor; aos meus amigos que em acompanharam desde a graduação e quem sempre tornaram os dias de trabalho duro mais suportáveis.

Agradeço à minha namorada, Mariana, pela compreensão, apoio, incentivo, carinho e por todo o amor que compartilhamos. Te amo.

E finalmente agradeço a todas as pessoas que de alguma forma me trouxeram inspiração: Se você realmente quiser alguma coisa, você conseguirá. São palavras como essas que tentarei levar comigo pelo resto da vida.

A todos vocês, muito obrigado!

Resumo

Com a grande proliferação de vídeos compartilhados na internet e o crescimento na sua utilização, cada vez mais torna-se indispensável a utilização de métodos automatizados para agrupar, analisar, indexar e buscar esses vídeos. Um dos tipos de análise de grande interesse atualmente é a análise semântica de vídeos de esportes devido as grandes possibilidades de aplicação comercial. Devido a grande diferença entre as regras e dinâmica de jogo, a abordagem mais comumente utilizada é primeiro realizar a identificação do gênero esportivo do vídeo para só então realizar uma análise semântica. Este processo é conhecido como categorização ou classificação de vídeos de esportes. A maior parte dos bancos de vídeos de esportes disponíveis para análise são compostos apenas por vídeos produzidos e transmitidos pela televisão. Neste trabalho, analisamos diversas técnicas para a classificação de vídeos de esportes e propomos uma combinação de características de cor (Autocorrelogramas) e de textura (*Local Binary Patterns - LBP*) para realizar a classificação do gênero esportivo em *frames* extraídos das sequências de vídeos. Nossa base de vídeos gerada para testes é composta por vídeos de três diferentes esportes, obtidos de fontes de diferente natureza: Vídeos capturados com equipamento profissional e transmitidos pela TV e sequências de vídeos geradas por usuários comuns através de *smartphones*. Esse tipo de tarefa representa um desafio porque vídeos amadores não são editados, as câmeras quase sempre se movem de maneira não-controlada e o ponto de visualização raramente é ideal. Nossa abordagem mostra uma taxa de classificação comparável com as técnicas do estado da arte quando as características são utilizadas separadas e um aprimoramento significativo quando são utilizadas de forma conjunta.

Palavras-chave: Autocorrelogramas. Padrões Binários Locais. Classificação de Vídeos de Esportes. Vídeos de Dispositivos Móveis. Classificador Multi-Classe. Máquina de Vetores Suporte. Fusão de Características.

Abstract

Due to the wide proliferation of videos shared on the internet and the increase in their use, automated methods to group, analyze, index and search these videos becomes indispensable. Currently, one type of analysis of great interest is the semantic analysis of sports videos due to the great potential for commercial application. Due to the big variance between the rules and game dynamics, one of the most commonly used solution is to first perform the identification of the genre of a sport video and only then perform a semantic analysis. This process is known as sports videos categorization or classification. Most sports videos databases available for analysis are composed only of videos produced and broadcasted by television. In this work we analyze several techniques for sports videos classification and propose combining features of color (Autocorrelograms) and texture (Local Binary Patterns - LBP) to perform classification of the sports genre within the frames extracted from video sequences. Our generate video database for testing consists of three different sports videos obtained from sources of different nature: Profissional produced videos from TV broadcast and video sequences generated by smartphone users. This kind of task is challenging because amateur videos are not edited, the cameras almost always move in uncontrolled manner and the point of view is rarely optimal. Our approach shows a correct classification rate comparable to the state of the art when features are used separated and significant improvement when the features combination is performed.

Keywords: Autocorrelograms. Local Binary Patterns. Sports Videos Classification. Mobile Videos. Multiclass Classifier. Support Vector Machine. Feature Fusion.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

AANN	Autoassociative Neural Network
AHC	Agglomerative Hierarchical Clustering
BOW	Bag of Words
CFS	Correlation-based Feature Selection
EDH	Edge Direction Histogram
EIH	Edge Intensity Histogram
GMM	Gaussian Mixture Models
GVP	Geometry-preserving Visual Phrases
HMM	Hidden Markov Model
KNN	K Nearest Neighbor
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
MFCC	Mel Frequency Cepstral Coefficient
MPEG	Moving Picture Experts Group
P2D-HMM	Pseudo-2D-Hidden Markov Model
PCA	Principal Components Analysis
PLSA	Probabilistic Latent Semantic Analysis
PNN	Probabilistic Neural Network
RNA	Rede Neural Artificial
RP	Random Projection
STIP	Space-time Interest Points
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
TF-RNF	Term Frequency - Relevance and Non-relevance Frequency

Sumário

1

Introdução

Atualmente vivendo na era da informação, estamos cercados por uma escala enorme de conteúdo digital. A estimativa é que o tráfego global de dados na internet ultrapassem 1 zettabyte por ano ao final de 2016 e que ainda em 2015 o compartilhamento de vídeos na internet correspondam a 61% de todos os dados trafegados (HESSELDAHL, 2014). Em termos de conteúdo de imagem e vídeo, o YouTube, o mais popular engenho de busca e compartilhamento de vídeos da atualidade, hospeda cerca 14 bilhões de vídeos, e novos *uploads* de vídeos são feitos à uma taxa de 100 horas por minuto (YOUTUBE, 2014).

Com esse volume de dados é impossível encontrar um vídeo de interesse vasculhando toda a base disponível o que torna indispensável o afunilamento de opções através da associação dos vídeos com categorias ou gêneros. A classificação automática de vídeos é uma área de pequisa que vem crescendo rapidamente nos últimos anos, especialmente devido a popularização do compartilhamento de conteúdo digital de programas de TV e a produção de vídeos pessoais com utilização de câmeras de baixo custo e *smartphones*.

É importante distinguir a classificação de vídeos do processo de indexação de vídeos, embora as escolhas de características e abordagens adotadas por ambas sejam similares. Muito da pesquisa de indexação é abordada da perspectiva de um banco de dados ser capaz de recuperar vídeos com eficiência e precisão, correspondendo a uma consulta realizada pelo usuário (HU et al., 2011). Por sua vez, algoritmos de classificação de vídeos colocam todos os vídeos em categorias, tipicamente um rótulo ou classe significativa (e.g., "vídeo de esporte"ou "vídeo de comédia"). Em geral, um sistema de classificação é composto das três etapas seguintes:

- **Extração de características:** Características são extraídos dos vídeos com técnicas de processamento de imagem e áudio.
- **Treinamento de um modelo de aprendizagem supervisionado:** Os vetores de características dos vídeos são utilizados para extrair conhecimento, para que um classificador possa "reconhecer similaridades"entre vídeos de uma mesma categoria.
- **Classificação:** Esta etapa consiste na aplicação do conhecimento extraído na etapa anterior. O classificador alimentado com os vetores de treinamento classifica vetores

extraídos de dados de teste não categorizados.

Categorização ou classificação de vídeos quanto ao gênero foi um dos primeiros tópicos em análise de vídeos a atrair o interesse dos pesquisadores. A principal atividade dessa categorização de gênero iniciou-se em grande grupos de vídeos, tais como esportes, música, notícias, filmes, etc., e gradualmente caminha para uma categorização mais refinada, tal qual a identificação de tipos de esportes.

A classificação de vídeos de esportes vêm mostrando uma série de aplicações comerciais (WANG et al., 2004), algumas sumarizadas na Figura 1.1. Grande parte da pesquisa na área trabalha na análise semântica desses tipos de vídeos em diferentes aplicações, tais como a combinação de vídeos de um mesmo evento esportivo obtido de diferentes usuários para produzir resumos de uma partida (melhores momentos) de forma automática. Os vídeos podem ser analisados para detectar *highlights* de um evento baseado em características de áudio ou visuais (XIONG, 2005).

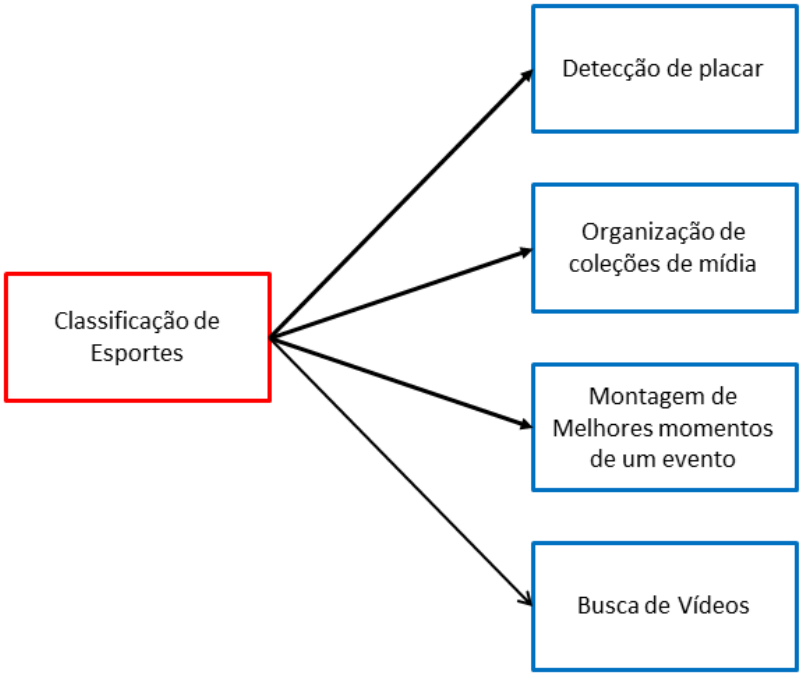


Figura 1.1: Aplicações da classificação de vídeo de esportes.

WANG et al. (2004) indica que a análise genérica baseada em conteúdo de vídeos de esportes é um desafio por conta da diversidade de regras nos esportes. A tarefa preliminar para tentar driblar a dificuldade de uma análise multimodal é primeiro categorizar os vídeos de acordo com o tipo de esporte analisado. Algumas aplicações da categorização ou classificação de esportes incluem a organização de coleções pessoais de mídia e busca de vídeos.

Segundo WU et al. (2009) vídeos considerados profissionais como o nome sugere são bem produzidos. Tais clipes são geralmente filmados em um ambiente controlado de estúdio com bons equipamentos de câmera, e são editados dependendo do gênero e conteúdo da história.

Por outro lado, vídeos amadores são filmados por hobby ou de maneira casual para capturar momentos interessantes. Tais clipes tem pouca ou nenhuma edição.

Neste trabalho , propomos uma abordagem robusta para categorizar de forma automática o gênero esportivo de uma base de dados misturando vídeos gravados de *smartphones* e vídeos de esportes produzidos de forma profissional para transmissão na TV. A base de vídeos utilizada é composta por vídeos de três diferentes esportes: futebol, vôlei e tênis.

Em particular nossa abordagem foca em distinguir os esportes uns dos outros utilizando apenas características visuais. As características utilizadas estão relacionadas a cor (Autocorrelogramas) e textura (*Local Binary Patterns* - LBP) das imagens. Máquinas de vetores suportes (*Support Vector Machines - SVM*) são utilizadas para classificar as características dos vídeos de forma individual, e posteriormente realizamos uma análise com duas abordagens de combinação de características. Esquemas de *early fusion* e *late fusion* são utilizados para avaliar se as características trabalham melhores juntas ou separadas.

Até então, este trabalho é o primeiro a tentar classificar o gênero esportivo em vídeos obtidos da transmissão de TV e gravados por usuários de *smartphones* ao mesmo tempo em uma base misturada. O desafio em classificar esse tipo de vídeo é trabalhar sem conhecimento *a priori* de características da natureza de produção dos vídeos tais como continuidade visual, comprimento da gravação e movimentação da câmera como levantadas em (NIU; LIU, 2012). Analisamos esse problemas avaliando as vantagens e inconvenientes de utilizar características visuais de processamento de sinal de baixo nível para a classificação de esportes.

1.1 Objetivos

Este trabalho define o seguinte conjunto de objetivos (não sendo a ordem um fator de importância):

- Realizar um estudo sobre o estado da arte da classificação de gêneros esportivos em vídeos de diferentes naturezas de produção.
- Propor e implementar uma abordagem para classificação de vídeos de esportes utilizando uma base de dados com vídeos amadores e profissionais.
- Realizar experimentos e validar a abordagem proposta de forma individual e completa (utilizando o sistema implementado), fazendo uma comparação com resultados de outras técnicas do estado da arte.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte maneira:

- No Capítulo 2 é feito um levantamento do estado da arte de classificação de vídeos de esportes com foco na escolha das características utilizadas para a realização dessa tarefa.

- O Capítulo 3 apresenta a técnica implementada para nossa abordagem de classificação utilizando um extrator de características de autocorrelogramas, o extrator LBP e o classificador SVM.

- No Capítulo 4 são descritos os experimentos realizados com a abordagem de combinação de características e o sistema proposto, bem como apresenta uma análise e discussão dos resultados. Também é descrita a construção da base de vídeos utilizada nos experimentos.

- No Capítulo 5 são apresentadas as conclusões do trabalho, contribuições e recomendações de trabalhos futuros.

2

Estado da Arte

Muitas abordagens já foram utilizadas para realizar classificação de vídeos em geral. BREZEALE; COOK (2008) dividiu essas abordagens em quatro grupos: abordagens baseadas em texto, abordagens baseadas em áudio, abordagens baseadas em vídeo e aquelas que utilizam alguma combinação de características visuais, de áudio ou textuais.

Classificação de vídeos pode ser realizada considerando vários aspectos, tais como cena "dentro de ambiente fechado"(*indoor*) ou "ao ar livre"(*outdoor*), "paisagem urbana"vs. "natureza", gênero, etc. Gêneros típicos são filmes, comerciais, notícias, esportes, música. Indo além, cada gênero pode ser ainda mais classificado em sub-gêneros (YUAN et al., 2006). O gênero filme por exemplo, pode ser subdividido em "ação", "drama", "comédia", etc., e o gênero esporte em "futebol", "basquete", etc. Nossa pesquisa foca em sub-gêneros de esportes e por questões de simplicidade nos referimos a eles apenas como "gêneros de esportes". Este capítulo apresenta uma revisão do estado da arte das principais características e modelos de aprendizagem utilizadas em classificação de gêneros de esportes.

Na primeira seção descrevemos trabalhos que abordam a classificação de vídeos de esportes mas consideram apenas conteúdo gravado por produções profissionais de TV. Muitos desses métodos obtém excelentes taxas de classificação aplicando suas abordagens sobre vídeos de produção profissional, mas devido a natureza das características utilizadas apresentariam uma queda de desempenho quando aplicados em vídeos amadores.

Na segunda seção, consideramos conteúdo não proveniente da TV, o qual é diferenciado por duas razões principais: é espontâneo e em sua maior parte não editado (consistindo de sequências cruas de vídeo), como apontado por GUO et al. (2012). Essa diferença possui um impacto direto no tipo de análise que pode ser realizada. Em vídeos amadores as câmeras são movidas de maneira descontrolada por usuários não profissionais. Em vídeos não editados não é possível explorar a presença de certos padrões de edição (como *replays*, placares, sequências de tamanho específico, duração de sequências antes da mudança de câmera, tipos de transição de câmera).

Vários métodos de fusão de informação foram propostos para agregar as predições de múltiplas fontes de informação em uma decisão global aperfeiçoada. As duas abordagens mais

comuns de combinar informações em tarefas de classificação são chamadas *early fusion* e *late fusion* (SNOEK; WORRING; SMEULDERS, 2005), as quais objetivam realizar a fusão nas etapa de pré-classificação e pós-classificação, respectivamente.

Finalmente na terceira seção, seguindo a metodologia de BREZEALE; COOK (2008) para realizar a revisão de literatura, fornecemos um panorama geral dos trabalhos publicados nessa área.

2.1 Classificação de Esportes em Vídeos Profissionais

Agrupamos os trabalhos de acordo com a utilização de características predominantes e diferenciadora da pesquisa. As imagens utilizadas para realizar os experimentos nesses trabalhos foram todas obtidas de transmissões de TV.

2.1.1 Características Baseadas em Cor

MUTCHIMA; SANGUANSAT (2012) propõe um esquema de ponderação de termos (*Term Weighting*) utilizado em pesquisas relacionadas a classificação de documentos para realizar categorização de vídeos de esportes. A ponderação de termos é aplicada para medir a importância de uma palavra para a classificação de conteúdo. A função que realiza tal tarefa depende da frequência de termos e da coleção de frequência das características. A abordagem proposta representa cada sequência de vídeo com *frames* que são extraídos de maneira uniforme. Cada *frame* de amostra é representado pelo seu histograma de cor para reduzir o tamanho da característica. Além disso, para incorporar alguma informação espacial o *frame* é particionado em quadrantes que possuirão seu próprios histogramas. A abordagem busca aplicar a técnica de projeção randômica (*Random Projection - RP*) para reduzir as dimensões de arquivos de vídeo com tamanho muito grande e propõem uma técnica chamada ponderação TF-RNF (*Term Frequency - Relevance and Non-relevance Frequency* weighting) cujo objetivo é filtrar e excluir os conteúdos não relevantes de forma a reduzir de maneira eficiente os erros de classificação.

Em WATCHARAPINCHAI et al. (2007) é introduzido uma característica de cor derivada dos correlogramas chamada autocorrelogramas que representa a correlação espacial entre cores idênticas da imagem para a classificação de vídeos de esportes. Essa característica é muito utiliza em problemas de indexação e busca de imagens. Os autores propõem a utilização dessa única características para discriminar seis esportes: boxe tailandês, futebol, golfe, salto ornamental e tênis. Os experimentos enfatizam a classificação de vídeo a nível de *frames* utilizando dois modelos de aprendizagem: Redes Neurais Artificiais com pré-processamento da técnica de redução PCA e SVM. Nesse estudo é utilizado um conjunto de dados de teste maior que o de dados de treinamento.

2.1.2 Características Baseadas em Movimento

Em GIBERT; LI; DOERMANN (2003) movimento e cor são utilizados para classificar vídeos de esportes em quatro tipos de classes: hóquei no gelo, basquete, futebol americano e futebol. Vetores de movimento de vídeo clipes MPEG são utilizados para associar um símbolo de direção de movimento para cada frame do vídeo. Símbolos de cores também são associados a cada pixel de cada frame. Uma representação da cor mais presente é associado ao frame.

Ao contrário da maioria da aplicações que utilizam modelos ocultos de Markov (*Hidden Markov Model - HMM*) para classificar vídeos, os autores treinam dois HMMs para cada classe de vídeo: Um para as representações de cores dos frames e outro para as representações de direção do movimento. A probabilidade de saída para cada classe é calculada pelo produto da probabilidade de saída de cor e movimento para aquela classe.

Em WANG et al. (2004) são propostos dois tipos de características. O primeiro tipo é a taxa de cor do campo ou quadra, mapeando as cores dos *pixels* para níveis pré-conhecidos de amarelo, verde e branco que por conhecimento de domínio dos autores são as cores mais comuns no campo de jogo dos esportes avaliados. A segunda é o movimento do plano de fundo (*background motion*) e taxa de movimento consistente (*consistency motion rate*). Esse último é a taxa de macroblocos internos cujo movimento é consistente com o plano de fundo. Essas características são computadas uma vez a cada quatro frames. Baseado nelas, um vetor de características 11-dimensional é calculado para cada sequência. Uma árvore de decisão C4.5 é então utilizada para classificar os esportes.

WANG; XU; CHNG (2006) tenta aumentar a performance da classificação utilizando características do domínio visual e de áudio assim como técnicas de análise multimodal demonstrando a efetividade do movimento global da câmera, cor dominante e características de *Mel-Frequency Cepstral Coefficients* (MFCC), utilizando uma *framework* de Pseudo-2D-HMM (P2DHMM) para modelar os padrões temporais em vídeos para o problema de classificação de vídeos de esportes. Esse modelo utiliza HMM, SVM e modelos de misturas Gaussiano (Gaussian Mixture Model - GMM) simultaneamente. O trabalho foca na utilização de campo de vetores de movimento (*Motion Vector Field*) disponível nos vídeos compactados para calcular a magnitude média do movimento (R^1), entropia do movimento (R^1), a direção dominante do movimento (R^1) e os fatores de *pan/tilt/zoom* (R^3) da câmera. *pan* e *tilt* referem-se a movimentos horizontais e de inclinação vertical da câmera respectivamente. TAN et al. (2000) descrevem bem o processo para obtenção dessas características. Além dessa característica a cor dominante é extraída com abordagens locais e global. A informação de áudio também está presente nessa abordagem por conta de certos sons específicos presentes nos esportes, e.g., o som do impacto da raquete com a bola no tênis, o atrito dos sapatos com a quadra no basquete, o apito em uma partida de futebol, etc. Como o som consiste principalmente de fala e som ambiente ou ruído, os MFCC utilizados em aplicações de reconhecimento de voz são extraídos. MFCCs de décima terceira ordem são extraído (R^{13}).

TAKAGI et al. (2003) buscou explorar a correlação entre o movimento da câmera e ações realizadas nos esportes, vídeos são categorizados com base em uma análise estatística dos parâmetros de movimento da câmera. Nesse método parâmetros do movimento da câmera são extraídos, vetores de movimento são classificados em oito direções e um histograma é calculado para cada categoria. Analisando as características do histograma, quatro tipos de movimentos de câmera são extraídos (*fix*, *pan*, *zoom*, *shake*) para cada grupo de imagens. No trabalho são introduzidas duas características:

- Taxa de extração do movimento da câmera (*Camera Motion Extraction Ratio*): representa a probabilidade de aparição de um dos parâmetros de movimento de câmera na sequência de vídeo. É definido como:

$$W_x = \frac{Num_{\text{aparição}}}{Num_{total}} \times 100(\%) \quad x \in fix, pan, zoom, shake \tag{2.1}$$

  onde Num_{total} é o número total de Grupos de imagens incluídas na sequência de vídeo e $Num_{\text{aparição}}$ é o número total de vezes que um parâmetro de movimento de câmera aparece.

- Transição de movimento de câmera (*Camera Motion Transition*): representa uma máquina de estados das transições dos parâmetros de movimentação de câmera:

$$T_{ab} = \frac{Num_{ab}}{Num_{\text{total-transições}}} \times 100(\%) \quad a, b \in x \tag{2.2}$$

  onde Num_{ab} é o número de transições de movimento de câmera entre o parâmetro A e B, e $Num_{\text{total-transições}}$ representa o total de transições de movimento de câmera na sequência de vídeo.

2.1.3 Características Baseadas em Borda

MOHAN; YEGNANARAYANA (2010) propõe a utilização de duas características baseadas em bordas: *Edge Direction Histogram - EDH* e *Edge Intensity Histogram - EIH*, as quais fornecem informação complementar para a classificação de vídeos de esportes. HMM e SVM são combinados nessa abordagem para classificar esportes em 5 categorias: basquete, críquete, futebol, tênis e vôlei.

- Histograma de direção de borda (*Edge Direction Histogram - EDH*) é um dos descritores visuais padrões de imagem e vídeo definidos no MPEG-7, e fornece uma boa representação de imagens com texturas não homogêneas. Esse descritor captura a distribuição espacial de bordas. A abordagem utilizada em MOHAN; YEGNANARAYANA (2010) é primeiro segmentar uma dada imagem em quatro

sub-imagens. A informação de borda é calculada para cada sub-imagem utilizando o algoritmo de Canny CANNY (1986). O intervalo de direções de borda $(0^\circ - 180^\circ)$ é quantizado em 5 *bins*. Assim, uma imagem particionada em 4 sub-imagens resulta em um vetor de características *EDH* com 20 dimensões.

- O histograma de intensidade de borda *Edge Intensity Histogram - EIH* é derivado da informação de magnitude dos pixels de borda. O intervalo de magnitude $(0-255)$ é quantizado em 16 pedaços, e um *EIH* com 16 dimensões é então derivado de cada frame de um vídeo. Por fim a classificação é realizada utilizando Redes Neurais Auto-Associativas (*Autoassociative Neural Networks - **ANN!***), HMM e SVM.

YUAN; WAN (2004) apresenta outra característica baseada em bordas, a percentagem de *pixels* nas bordas, que é extraída de cada frame representativo para classificar um dado vídeo de esporte e uma de cinco categorias: *badminton*, futebol, basquete, tênis e patinação artística. Para cada frame representativo a imagem de bordas é obtida com detector de bordas de Canny. As imagens são então transformadas em imagens dilatadas. Para uma imagem de altura H e comprimento W a imagem é dividida em blocos de tamanho NxN. Para cada bloco, se mais de M *pixels* de borda forem encontrados, o bloco é identificado como bloco de borda e todos os seus pixels passam a ter valor 1, caso contrário o bloco é considerado um bloco de não-borda e seus *pixels* passam a ter valor zero. A taxa de borda (*edge ratio*) de uma imagem é definida como:

$$\text{taxa de borda} = \frac{S_b}{S} \tag{2.3}$$

onde S_b é a soma da quantidade de *pixels* de borda e S a quantidade de *pixels*. Posteriormente a taxa de borda média é calculada para uma sequência de vídeo, realizando a média dos *frames* amostrados nessa sequência. O algoritmo KNN foi utilizado para a classificação desses vídeos.

2.1.4 Características em Técnicas de Produção do Gênero

Nos referimos a "técnicas de produção de gênero"como elementos de filmagem utilizadas em esportes. A transmissão de jogos pela TV alterna entre sequências, filmados por diferentes câmeras para dar maior dinamismo aquela partida.

O método proposto em SIGARI; SURESHJANI; SOLTANIAN-ZADEH (2011) utiliza seis características: três cores dominantes (*dominant color*) , o nível de cinza dominante (*dominant grey level*), a taxa de corte (*cut rate*) e taxa de movimento (*motion rate*). As cores dominantes são obtidas no espaço HSV através de histogramas com 32 bins. A média dos histogramas de todos os frames é calculada e então as três cores dominantes são extraídas do histograma médio. O nível dominante de cinza é utilizado para discriminar esportes cuja cor dominante é cinza como esqui e hóquei no gelo. A transição entre sequências de gravação é dividida em duas categorias: **abrupta ou transição com corte** e **transição gradual**. Segundo

os autores a taxa de corte que é obtida pelo método em AMIRI; FATHY (2009) fornece uma medida do excitamento no vídeo e pode ser utilizado porque diferentes esportes possuem diferentes padrões de excitamento. A taxa de movimento é obtida através da média das diferenças bit a bit em frames consecutivos de cada vídeo. Essas características são fornecidas para um classificadores em um comitê (*ensemble*): Nearest Neighbor (NN), Linear Discriminant Analysis (LDA), Árvore de Decisão e uma Rede Neural Probabilística (*Probabilistic Neural Network - PNN*). O voto de cada um ponderado pela taxa de classificação fornece a decisão final.

2.1.5 Características Implementadas no Padrão MPEG-7

A pesquisa em XU et al. (2008) utiliza quatro algoritmos da área de aprendizagem de máquina: Árvores de Decisão, Support Vector Machines (SVM), *K-Nearest Neighbor* e *Naive Bayesian*, para avaliar o impacto em performance de técnicas de redução de dados na categorização de vídeos de esportes. As técnicas de redução utilizadas são *Correlation-based Feature Selection* (CFS), Principal Component Analysis (PCA) e *Relief*. As característica utilizadas fazem parte do padrão MPEG-7 e são brevemente descritos a seguir:

- *Color-Structure Descriptor* é uma característica que captura conteúdo de cor (similar a um histograma) e informação sobre a estrutura desse conteúdo.
- *Color Layout* específica a distribuição espacial de cores.
- *Edge Histogram* o histograma de borda representa a distribuição espacial de cinco tipos de bordas: quatro bordas direcionais e uma borda não direcional.
- *Region-Based Shape* faz uso de todos os pixels que constituem uma forma dentro de um frame, para descrever uma região ou conjunto de regiões.
- *Homogenous Texture Descriptors* uma imagem pode ser considerada como um mosaico de texturas homogêneas então essas características de texturas são utilizadas para indexar dados de imagens.
- *Texture Browsing* requer apenas 12 bits e fornece uma caracterização de percepção semelhante à humana, em termos de regularidade, rudeza e direcionalidade.

2.1.6 Outros Tipos de Características

LI et al. (2009) utilizam uma base de vídeos de larga escala e realizam a classificação utilizando características SURF (*Speeded-up Robust Features*) de imagens amostradas de maneira uniforme de sequência decodificadas de frames. Após obter as características SURF do vídeo de treinamento, um *codebook* é gerado utilizando o algoritmo de clusterização k-means e cada valor de *codeword* é definido pelo vetor de exemplo de cada cluster. Mapeando cada descritor SURF individual de uma frame para um *codebook*, cada frame pode ser representado por um

histograma que mostra a frequência de aparição de cada *codeword* no *codebook*. Um esquema de categorização hierárquica é utilizado, separando os esportes em categorias semânticas, agrupando por exemplo futebol, basquete e vôlei num grupo de esportes maior: esportes com bola. A divergência de Kullback-Leibler (KL) proveniente da teoria da informação e mostrada na equação 2.4, é utilizada pelos autores para medir a distância entre a distribuição Q que representa um vídeo de teste e a distribuição T de uma categoria de treino. Após isso o algoritmo KNN é utilizado para categorizar o gênero do esporte baseado na divergência KL.

$$D_{KL}(Q\|T) - \sum_i [q_i \cdot log(q_i/t_i)] \tag{2.4}$$

Em YUAN et al. (2006) é apresentado uma esquema de categorização de vídeos baseado em uma ontologia de gêneros hierárquica. Dez características espaço-temporais são extraídas para distinguir diferentes gêneros utilizando um SVM hierárquico que consiste de uma série de SVMs unidos na forma de uma árvore binária. O trabalho primeiro classifica os gêneros em filme, comercial, notícias, esportes e vídeos de música e em seguida classifica filmes e esportes em sub-gêneros. No entanto o nosso interesse é na classificação de vídeos de esportes e as categorias classificadas são: futebol, basebol, basquete, futebol americano, tênis e vôlei. As características estão divididas em dois grupos:

- *Temporais*: incluem tempo médio de gravação de uma sequência, taxa de corte (*cut percentage*), diferença de cor média e movimento de câmera. A diferença de cor média é calculada realizando a média entre todos os histogramas de cor de todo o vídeo. A diferença de dois histogramas de cor consecutivos é dada por:

$$1 - \frac{\sum_{i=1}^{N} \min(H_j(i), H_{j-1}(i))}{\sum_{i=1}^{N} H_j(i)} \tag{2.5}$$

  onde N é numero de *bins* do histograma de cor, e H_j e H_{j-1} são os histogramas de cor do frame j e $j-1$, respectivamente.

- *Espaciais*: São a taxa de frames de face (*frames face ratio*), brilho médio (*average brightness*) e entropia de cor média (*average color entropy*).

Dois tipos de formas de árvores binárias com SVM são aplicadas: Ótimo local que busca a separação de cada gênero/nó da hierarquia de forma local e o ótimo global que busca a melhor separação de todos os nós da árvores de uma só vez.

Finalmente em DONG et al. (2012) é apresentada uma abordagem baseada em extração de sequências representativas e frases visuais de geometria (*Geometry-preserving Visual Phrase - GVP*) para categorização de esportes. *Frames* representativos são mesclados em vários *clusters* utilizando um método de clusterização aglomerativa hierárquica (*Agglomerative Hierarchical Clustering - AHC*) para que através de regras pré-definidas o mais representativo seja escolhido.

GVPs, descritas em ZHANG; JIA; CHEN (2011), são buscada nas sequências de vídeo selecionadas e frames são representados como um histograma de palavras visuais (*visual words*) e frases visuais (*visual phrases*). Classificadores SVM utilizando múltiplos tipos de kernels são então utilizados para classificar os gêneros dos esportes.

2.2 Classificação em Vídeos Amadores

Dividimos essa seção em duas partes: pesquisas que classificam vídeos produzidos por "usuários comuns"e uma outra seção para agrupar imagens que não são transmitidas pela televisão mas foram geradas de maneira controlada. Vídeos produzidos por usuários são aqueles gravados por câmeras de baixo custo ou celulares sem o intuito específico de ser utilizado para fins de pesquisa.

2.2.1 Classificação de Vídeos Gerados Por Usuários

CRICRI et al. (2013) apresentam uma abordagem para classificar de forma automática vídeos de seis esportes gerados por usuários, gravados no mesmo evento esportivo utilizando dispositivos móveis. O trabalho analisa múltiplas modalidades de dados capturadas por diferentes usuários. divididas em três categorias:

- Características Visuais: utilizam os descritores do padrão MPEG-7, *dominant color*, *color layout*, *color structure*, *edge histogram*, textithomogenous texture. O vetor de característica visuais é formado concatenando cada uma dessas características.
- Características de Áudio: são extraídas as características de coeficientes MFCC de áudio monofônico com taxa de amostragem de 44,1 KHz. Vetores de características com 12 coeficientes MFCC são obtidos de janelas de 40ms sem sobreposição.
- Características de sensores auxiliares: Os sensores auxiliares são uma bússola eletrônica e um acelerômetro (dispositivos encontrados na maioria dos smartphones) e são utilizados através de software dedicado para modelar o movimento da câmera.

Os autores propõem adaptar dinamicamente a fusão de diferentes modalidades baseadas na qualidade apresentada para cada vídeo de entrada, utilizando as características visuais e de sensores auxiliares para treinar SVMs e as características de áudio para treinar Modelos de Mistura Gaussiana (GMM).

O trabalho é estendido em CRICRI et al. (2014) para analisar a extração de características espaço-temporais que são menos afetadas pela movimentação da câmera e utilização de modelos *Bag-of-Visual-Words (BoW)* para representar os aspectos espaço-temporais dessas características e realizam uma extensiva análise de adaptação de fusão de características seguindo diferentes abordagens de *early fusion*, *intermediate fusion* e *late fusion*. Um modelo para *late fusion* chamado 2D *majority voting* também é proposto.

2.2.2 Classificação de Vídeos Controlados

A ideia principal no método apresentado por LEE; HOFF (2007) é se uma trajetória é discretizada e representada como uma sequência de símbolos, cada um deles representando um tipo de comportamento atômico, o comportamento coletivo de um tipo de atividade, ou a assinatura do tipo de atividade, pode ser modelada por uma cadeia de Markov no tempo discreto. Para n atividades de interesse são coletadas trajetórias dessas atividades que são posteriormente segmentadas e discretizadas e n modelos de Markov são criados, 1 para cada tipo de atividade. A trajetória pertencerá para o tipo de atividade que resultar na maior probabilidade. As trajetórias são extraídas rastreando os objetos que se movem no vídeo e são divididas em conjuntos de teste e treinamento. Essa técnica é utilizada para classificar vídeos quanto a trajetória detectada de jogadores separando-os em duas classe: *frisbee* e vôlei.

Em GADE; MOESLUND (2013) é apresentado um foco diferente dos demais trabalhos na área, realizando a classificação de vídeo para identificar dentro de uma mesma quadra, quais esportes estão sendo jogados a fim de fornecer informações para otimização de uso da quadra e suas dependências. Para realizar a categorização de esportes são utilizadas apenas informações baseadas em posição, através de mapas de ocupação de calor (*heatmaps*) obtidos com uma câmera especial que produz imagens térmicas. Os jogadores na quadra e suas posições são detectados utilizando homografia. Os mapas de calor são produzidos pela sumarização de distribuições gaussianas e antes da classificação são projetados para um espaço discriminativo de dimensões reduzidas pelo princípio de *Fischerfaces* BELHUMEUR; HESPANHA; KRIEGMAN (1997), uma combinação da utilização de PCA e Discriminante Linear de Fischer, utilizado em reconhecimento de face.

2.3 Panorama de Classificação de Vídeos de Esportes

A Tabela 2.1 apresenta um sumário das principais técnicas de classificação de vídeos de esportes apresentadas neste capítulo. A primeira coluna apresenta as principais referências apresentadas durante a revisão de literatura, embora algumas tenham sido omitidas por não estarem diretamente relacionadas ao nosso objetivo de classificação em vídeos de esportes profissionais e amadores. A segunda coluna da tabela descreve os esportes discriminados pelas técnicas informando também a quantidade. Na terceira coluna e quarta coluna são apresentadas as características e modelos de aprendizagem utilizados. No capítulo 4 é apresentado um complemento dessa tabela informando o tamanho das bases de vídeo e taxas de classificação obtidas.

Tabela 2.1: Resumo das informações sobre pesquisas da revisão de literatura.

Referência	**Esportes**	**Características**	**Modelos para Categorização**
YUAN et al. (2006)	basebol, basquete, futebol, futebol americano, tênis e vôlei (6)	*shot length, cut percentage, average color difference, camera motion, face frames ratio, average brightness e average color entropy*	SVM hierárquico
DONG et al. (2012)	basquete, ciclismo, boxe, atletismo, handebol, formula-1, rúgbi , patinação, futebol e tênis (10)	GVPs	SVM
GIBERT; LI; DOERMANN (2003)	basquete, futebol, futebol americano e hóquei no gelo (4)	*dominant color, motion direction*	HMMs
MOHAN; YEGNANARAYANA (2010)	basquete, críquete, futebol americano, tênis e vôlei (5)	*edge direction histogram, edge intensity histogram*	AANN, HMM, SVM
LEE; HOFF (2007)	frisbee e vôlei (2)	*short trajectories*	Modelos de Markov
LI et al. (2009)	basquete, motociclismo, boxe, natação, mergulho, judô, formula-1, patinação artística, patinação de velocidade, futebol, sinuca, tênis, tênis de mesa e vôlei (14)	BoW de descritores SURF	SVM, PLSA

MUTCHIMA; SANGUANSAT (2012)	basquete, vôlei de praia, corrida de bicicletas, *lawn bowls*, boliche, boxe, corrida de carros, futebol americano, hóquei, corrida de motocicletas, rúgbi, esqui, sinuca, *squash*, natação, tênis de mesa, tênis, vôlei, *walkathon* e luta livre (20)	histograma de cor	KNN
WANG et al. (2004)	basebol, basquete, hóquei no gelo e golfe (4)	*color ratio, background motion, consistency motion ratio*	Árvore de Decisão C4.5
WANG; XU; CHNG (2006)	basquete, futebol e tênis (3)	*average motion magnitude, motion entropy, dominant motion direction, camera pan/tile/zoom factors, local/global dominant colors, MFCC*	Pseudo-2D-HMM
SIGARI; SURESHJANI; SOLTANIAN-ZADEH (2011)	futebol americano, basquete, tênis, natação, futebol de salão, esqui e boxe (7)	*3 dominant colors, dominant gray level, cut rate and motion rate*	comitê de classificadores: NN, LDA, Árvore de Decisão, PNN
WATCHARAPINCHAI et al. (2007)	basquete, boxe tailandês, futebol americano, golfe, mergulho, tênis e vôlei (7)	autocorrelogramas	RNA, SVM

XU et al. (2008)	basquete, futebol, tênis de mesa e natação (4)	*color-structure descriptor, color layout, edge histogram, region-based shape, homogenous texture descriptors, texture browsing*	Árvore de Decisão, KNN, SVM, Navie Bayes
YUAN; WAN (2004)	*badminton*, basquete, patinação, futebol e tênis (5)	*average edge ratio*	KNN
CRICRI et al. (2013)	futebol, futebol americano, basquete, tênis, hóquei no gelo e vôlei (6)	*dominant color, color layout, color structure, scalable color, edge histogram, homogeneous texture*, MFCC, *auxiliary sensor features*	SVM com esquemas de *early, intermediate late fusion*
CRICRI et al. (2014)	futebol, futebol americano, basquete, tênis, hóquei no gelo e vôlei (6)	*dominant color, color layout, color structure, scalable color, edge histogram, homogeneous texture*, MFCC, *auxiliary sensor features*, *Bag of Words de características STIP*	SVM com ainda mais esquemas de *early, intermediate late fusion*

3

Combinação de Características e Classificação de Esportes

Para o problema de classificação de vídeos de esportes como mostrado no capítulo anterior uma série de características já foi utilizada para realizar o treinamento e teste de diferentes modelos de aprendizagem. Especificamente para nossa abordagem classificando vídeos amadores e profissionais optamos por utilizar apenas características visuais para evitar possíveis problemas gerados por ruído e a grande variação no áudio das gravações de esportes.

O trabalho apresentado em SUGANO et al. (2009) discute que devido à baixa qualidade de produção de vídeos amadores, as caraterísticas extraídas são menos confiáveis que aquelas extraídas de vídeos profissionais. Dessa forma características robustas ao movimento inconsistente da câmera precisam ser incorporadas na análise. Evitamos a utilização de características já implementadas no padrão MPEG porque isso restringe o formato dos vídeos passíveis de classificação pela técnica implementada e optamos por características comuns ao contexto de vídeos amadores e profissionais.

Este capítulo descreve a abordagem utilizando apenas características visuais para classificar corretamente vídeos amadores e profissionais de esportes, combinando as características a fim de obter uma melhor classificação utilizando SVM.

3.1 Autocorrelogramas

Vídeos de esportes gravados por amadores geralmente apresentam abruptos movimentos da câmera quando estão gravando um esporte com movimentação dinâmica como o futebol e estão fortemente relacionados com o posicionamento do espectador no estádio, campo ou quadra (CRICRI et al., 2013). Logo, a utilização de autocorrelogramas como característica para classificar esses tipos de vídeos é uma escolha apropriada porque tolera grandes mudanças causadas na aparência e forma causadas pela alteração do ponto de visualização da cena, zoom da câmera, etc. Além disso autocorrelogramas são características facilmente computáveis e o tamanho final da característica é consideravelmente pequeno (HUANG et al., 1997).

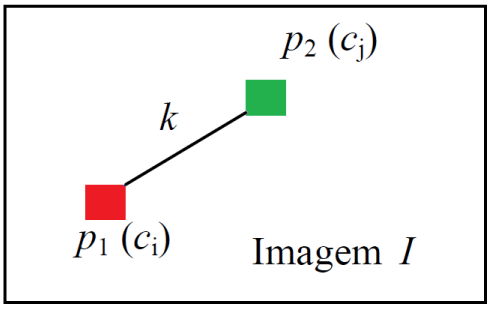


Figura 3.1: A figura ilustra o conceito de correlograma de um pixel p_1 de cor c_i e um pixel p_2 a uma distância k de p_1. O correlograma de cor é definido como a probabilidade de que o pixel p_2 seja da cor c_j. Adaptado de SKULSUJIRAPA; ARAMVITH; SIDDHICHAI (2004)

Um correlograma de cor expressa como a correlação espacial dos pares de cores muda com a distância, diferentemente de um histograma que captura apenas a distribuição de cor em uma imagem e não inclui qualquer informação espacial.

O conceito de correlograma de cor é ilustrado na Figura 3.1. Na imagem I, escolhendo qualquer pixel p_1 de cor c_i e outro pixel p_2 a uma distância k de p_1, o correlograma de cor é definido como a probabilidade de que o pixel p_2 seja da cor c_j.

Seja I uma imagem de tamanho $n_1 \times n_2$. As cores em I são quantizadas em m cores $c_1, c_2, \ldots, c_m$. Para os pixels $p_1 = (x_1, y_1)$ e $p_2 = (x_2, y_2)$, usamos a medida L_∞-norm para medir a distância entre os pixels. A distância L_∞-norm é definida como $|p_1 - p_2| = \max\{|x_1 - x_2|, |y_1 - y_2|\}$ e $k \in \{1, 2, \ldots, n\}$ é o conjunto de distâncias. Essa distância é utilizada em HUANG et al. (1997) com a justificativa de por conveniência ser um cálculo simples para medir distância entre pixels. A Figura 3.2 demarca os pixels para uma determinada distância k utilizando a medida L_∞-norm.

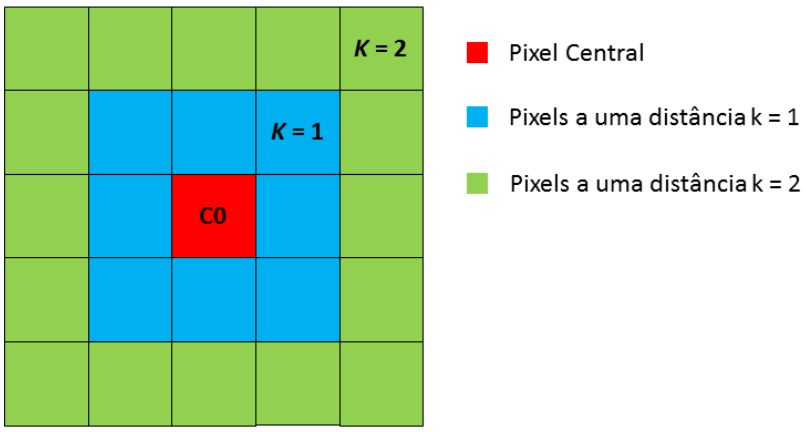


Figura 3.2: Pixels a uma distância L_∞-norm especificada. Repara-se que o número de pixels a uma distância k é sempre $8 \times k$.

Logo, para uma distância k fixa, o correlograma de cor dos pixels da imagem I é definido

como:

$$\gamma_{c_i,c_j}^{(k)} \triangleq \underset{p_1 \in I_{c_i}, p_2 \in I}{Pr}[p_2 \in I_{c_j} || p_1 - p_2| = k] \tag{3.1}$$

Onde I_{c_i} representa o subconjunto de pixels de I que são da cor c_i. Assim sendo, o autocorrelograma de I captura a correlação espacial apenas entre cores idênticas e é definido como:

$$\alpha_c^{(k)} \triangleq \gamma_{c,c}^{(k)} \tag{3.2}$$

Essa informação é um subconjunto do correlograma e requer apenas $m \times k$ dimensões. A complexidade do método de autocorrelogramas é diretamente relacionada à escolha da distância entre pixels (k) e o tamanho da imagem $n_1 \times n_2$. A escolha do parâmetro k para definir o correlograma está sujeita a seguinte questão: Um valor grande para k pode resultar em aumento em complexidade e custo computacional enquanto que um valor pequeno pode comprometer a qualidade da característica.

3.1.1 Computando Correlogramas

Para computar o correlograma, é suficiente realizar a contagem mostrada na Equação 3.3 (similar a matriz de co-ocorrência definida em HARALICK (1979) para análise de texturas de imagens em nível de cinza).

$$\Gamma_{c_i,c_j}^{(k)}(I) \triangleq |\{p_1 \in I_{c_i}, p_2 \in I_{c_j} || p_1 - p_2| = k\}| \tag{3.3}$$

Para $\gamma_{c_i,c_j}^{(k)}(I) = \Gamma_{c_i,c_j}^{(k)}(I)/(h_{c_i}(I) \times 8k)$, onde o termo $h_{c_i}(I)$ representa o histograma h de I e é definido na Equação 3.4:

$$h_{c_i}(I) \triangleq n^2 \cdot Pr_{p \in I}[p \in I_{c_i}] \tag{3.4}$$

Para qualquer pixel na imagem, $h_{c_i}(I)/n^2$ retorna a probabilidade que a cor do pixel seja c_i.

O denominador $(h_{c_i}(I) \times 8k)$ é o número total de pixels a uma distância k de qualquer pixel de cor c_i (O fator $8k$ é devido as propriedades da L_∞-norm como mostrado na Figura 3.2).

3.2 *Local Binary Patterns*

O *Local Binary Patterns* é um operador de imagem que transforma uma imagem em uma matriz ou imagem de rótulos com valores inteiros. Esses rótulos ou as estatísticas relacionadas a eles, mais comumente o histograma, são então utilizados para análise de textura da imagem. As versões mais amplamente utilizadas do operador são voltadas para imagens monocromáticas

(em nível de cinza) mas ele foi estendido também para imagens coloridas, assim como vídeos e dados volumétricos. No nosso trabalho utilizamos o LBP apenas no domínio espacial.

3.2.1 LBP Básico

O operador LBP básico original, introduzido por OJALA; PIETIKAINEN; HARWOOD (1996) foi baseado na ideia de que textura pode ser descrita por dois aspectos complementares: padrões espaciais locais e contraste do nível de cinza (intensidade do pixel). A versão original do operador LBP foi projetada para trabalhar com uma máscara de tamanho 3x3 pixels. Os pixels nessa máscara são limiarizados pelo valor do pixel no centro da máscara (o valor 1 é assinalado para valores intensidade maior ou igual ao do pixel central, caso contrário o valor 0 é assinalado), multiplicados por potências de dois e então somados para obter um rótulo para o pixel central. A Figura 3.3 exemplifica o processo.

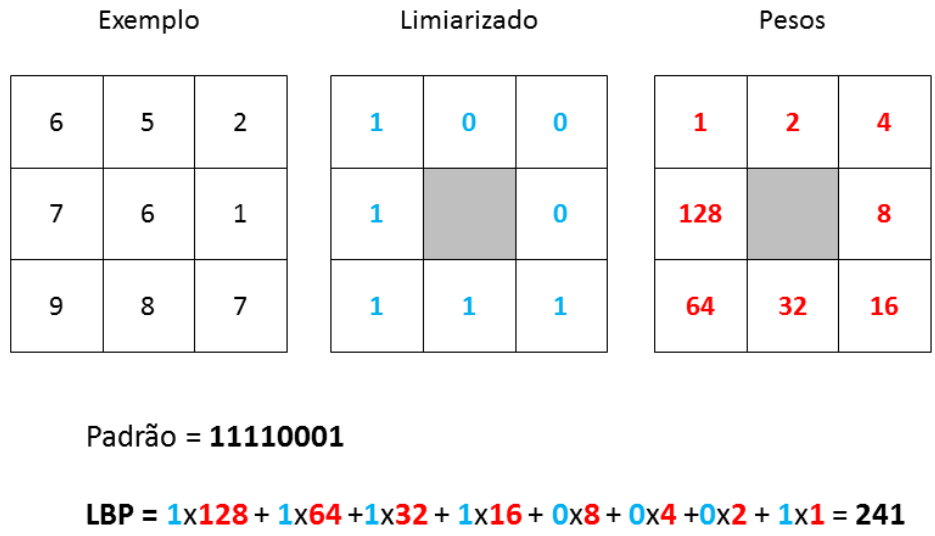


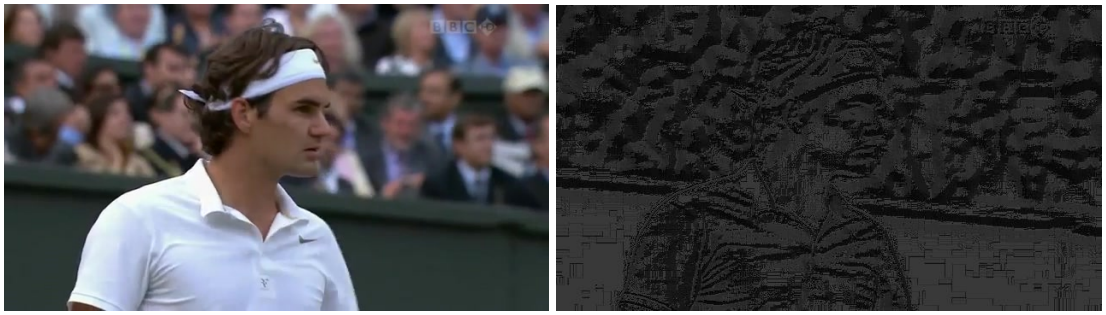
Figura 3.3: Cálculo do código LBP de uma vizinhança 3x3.

Como a vizinhança consiste de 8 pixels, um total de $2^8 = 256$ rótulos (códigos LBP binários) diferentes podem ser obtidos dependendo da relação entre os níveis de cinza do pixel central e os níveis de cinza da vizinhança.

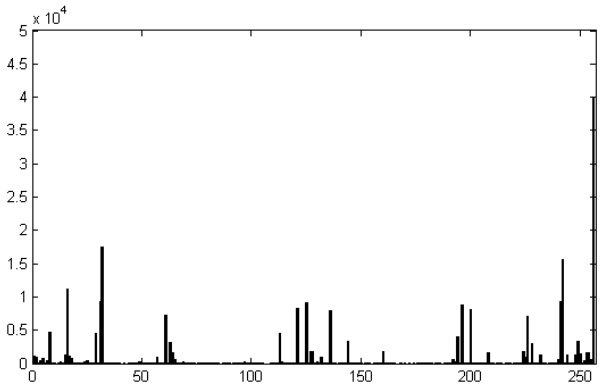
A Figura 3.4 mostra um exemplo de uma imagem LBP obtida a partir de um frame da nossa base de teste, juntamente com o seu histograma.

3.2.2 Operador LBP Genérico

Anos após a publicação de sua versão original, o operador LBP foi apresentado em uma forma revisada mais genérica em OJALA; PIETIKAINEN; MAENPAA (2002). Em contraste com o LBP original que utiliza 8 pixels em uma máscara 3x3, essa formulação genérica do



(a) Imagem original. (b) Imagem LBP.



(c) Histograma LBP.

Figura 3.4: Exemplo de uma imagem de entrada, a imagem LBP correspondente e seu histograma.

operador não impõe limite para o tamanho da vizinhança ou para o número de pontos de amostragem. A derivação do LBP genérico é apresentada a seguir.

Considere uma imagem monocromática I , a matriz de níveis de cinza dessa imagem $I(x,y)$ e seja g_c o nível de cinza de um ponto (pixel) arbitrário de I com coordenadas (x,y), i.e. $g_c = I(x,y)$. Além disso, g_p denota o nível de cinza de um ponto de amostragem em uma vizinhança circular com P amostras igualmente espaçadas em um raio R ao redor do ponto com coordenadas (x,y):

$$g_p = I(x_p, y_p), \quad p = 0, \ldots, P-1 \tag{3.5}$$

$$x_p = x + Rcos(2\pi p/P), \tag{3.6}$$

$$y_p = y - Rsin(2\pi p/P). \tag{3.7}$$

A Figura 3.5 mostra exemplo de vizinhanças circulares. Assumindo que a textura local da imagem $I(x,y)$ é caracterizada pela distribuição conjunta de níveis de cinza $P+1(P>0)$ pixels:

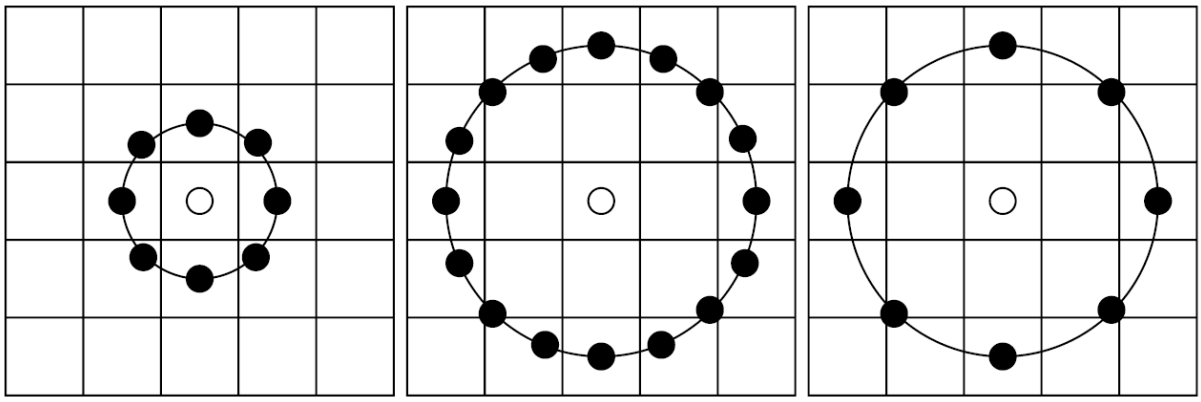


Figura 3.5: As vizinhanças circulares $(8,1)$, $(16,2)$ e $(8,2)$. Os valores dos *pixels* são interpolados de forma bilinear sempre que um ponto de amostragem não estiver no centro de um pixel. Adaptado de PIETIKAINEN et al. (2011).

$$T = t(g_c, g_0, g_1, \ldots, g_{P-1}) \tag{3.8}$$

Sem perda de informação, o valor do pixel central pode ser subtraído da vizinhança:

$$T = t(g_c, g_0 - g_c, g_1 - g_c, \ldots, g_{P-1} - g_c) \tag{3.9}$$

No próximo passo a distribuição conjunta é aproximada assumindo-se que o pixel central é estatisticamente independente das diferenças, o que permite a fatoração da distribuição:

$$T \approx t(g_c)t(g_0 - g_c, g_1 - g_c, \ldots, g_{P-1} - g_c) \tag{3.10}$$

Agora o primeiro fator $t(g_c)$ é a distribuição de intensidade sobre $I(x,y)$. Do ponto de vista da análise de padrões de textura locais, essa distribuição não contém informação útil. Ao invés disso a distribuição conjunta de diferenças

$$t(g_0 - g_c, g_1 - g_c, \ldots, g_{P-1} - g_c) \tag{3.11}$$

pode ser utilizada para modelar a textura local. No entanto, uma estimativa confiável dessa distribuição multidimensional de dados da imagem pode ser complicada. Uma solução para esse problema, proposta por OJALA et al. (2001), é aplicar quantização de vetores. Eles utilizaram a quantização de vetores com um *codebook* de 384 códigos para reduzir a dimensionalidade do espaço de características. Os índices dos 384 códigos correspondem a 384 bins (códigos LBP) no histograma. Assim, esse poderoso operador baseado em diferenças entre níveis de cinza pode ser considerado um operador texton. A abordagem baseada em LVQ (*Learning Vector Quantization*) ainda possui certas propriedades indesejáveis que tornam sua utilização difícil. Primeiramente, as diferenças $g_p - g_c$ são invariantes a mudanças do nível de cinza médio da imagem mas não são invariantes em relação a outras mudanças no nível de cinza. Para ser utilizado para classificação de textura o codebook deve ser treinado de forma similar aos outros

métodos baseados em textons. Para suavizar esses desafios, apenas os sinais das diferenças são considerados:

$$t(s(g_0 - g_c), s(g_1 - g_c), \ldots, s(g_{P-1} - g_c)), \tag{3.12}$$

onde $s(z)$ é a função de limiar definida como:

$$s(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$$

O operador LBP genérico é derivado dessa distribuição conjunta. Como no caso do LBP original, ele é obtido somando as diferenças limiarizadas ponderadas por potências de dois. O operador $LBP_{P,R}$ é definido como:

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p. \tag{3.13}$$

Na prática a Equação 3.13 significa que os sinais das diferenças em uma vizinhança são interpretados como um número binário de P bits, resultando em 2^P valores distintos para o código LBP. A distribuição local de nível de cinza, i.e. textura, pode ser então aproximadamente descrita com uma distribuição discreta com 2^P de códigos LBP:

$$T \approx t(LBP_{P,R}(x_c, y_c)). \tag{3.14}$$

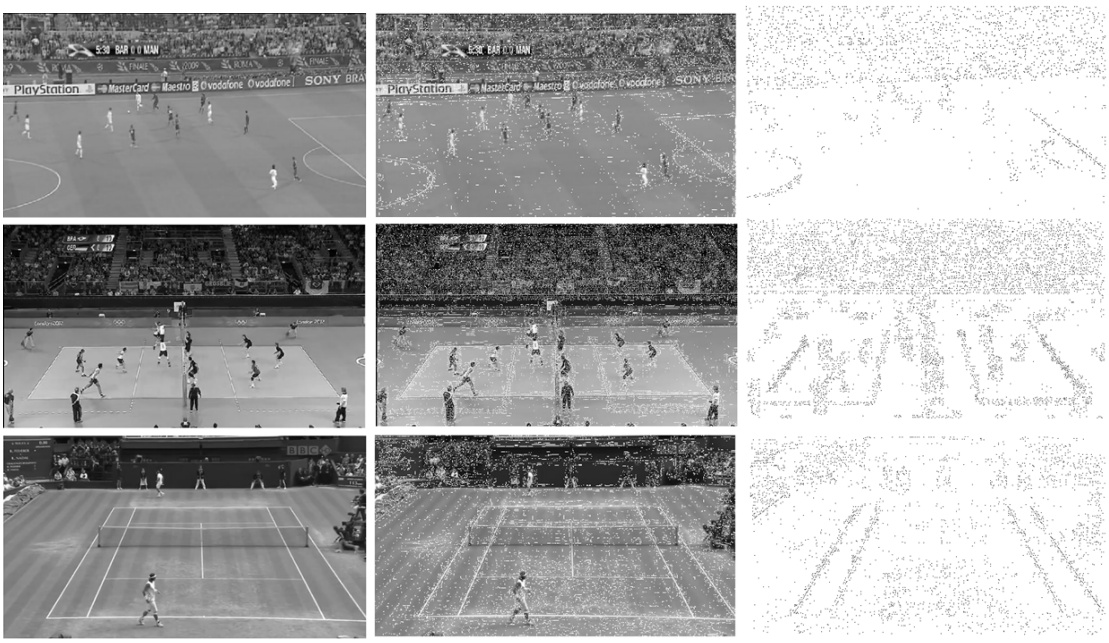
No cálculo da distribuição $LBP_{P,R}$ (vetor de características) para uma imagem de teste $N \times M$ com $x_c \in \{0, \ldots, N-1\}, y_c \in \{0, \ldots, M-1\}$, a parte central é considerada somente porque uma vizinhança suficientemente grande não pode ser utilizada nas bordas. O código LBP é calculado para cada pixel na porção cortada da imagem, e a distribuição dos códigos é utilizada como um vetor de características, denotado por S:

$$S = t(LBP_{P,R}(x,y)), x \in \{\lceil R \rceil, \ldots, N-1-\lceil R \rceil\}, y \in \{\lceil R \rceil, \ldots, M-1-\lceil R \rceil\}. \tag{3.15}$$

O LBP original é muito similar ao $LBP_{8,1}$, com duas diferenças. A primeira é que a vizinhança na definição geral é indexada de forma circular, tornando mais fácil de obter descritores de texturas invariantes a rotação. A segunda, que os pixels da diagonal na vizinhança 3x3 são interpolados no $LBP_{8,1}$.

3.2.3 Padrões Uniformes

Em muitas aplicações de análise de textura é desejável ter características que são invariantes ou ao menos robustas a rotações da imagem processada. Como os padrões $LBP_{P,R}$ são obtidos por amostragem circular ao redor do pixel central, a rotação da imagem de entrada tem



(a) Imagens originais. (b) Imagens destacando pixels com padrões uniformes. (c) Imagens destacando pixels com padrões não-uniformes.

Figura 3.6: Imagens mostrando a grande maioria de pixels com padrões uniformes. A cor dos pixels não destacados é substituída pela cor branca para facilitar a visualização.

dois efeitos: cada vizinhança local é rotacionada para outra localização de pixel, e dentro de cada vizinhança, os pontos de amostragem no círculo ao redor do ponto central são rotacionado numa orientação diferente.

Outra extensão do LBP original utiliza os chamados *padrões uniformes* (OJALA; PIETIKAINEN; MAENPAA, 2002). Para essa tarefa, uma métrica de uniformidade de padrão é utilizada: U é o número de transições bit a bit de 0 para 1 ou vice versa quando o padrão do bit é percorrido de maneira circular. Um código LBP é considerado uniforme se a medida de uniformidade U é no máximo 2. Para exemplificar, os padrões 00000000 (zero transições), 01110000 (duas transições) e 11001111 (duas transições) são uniformes enquanto os padrões 11001001 (quatro transições)e 01010010 (seis transições) não são. No mapeamento uniforme LBP é utilizado um rótulo binário cada padrão uniforme e todos os padrões não-uniformes são mapeados para um único rótulo. Assim, o número de rótulos diferentes para mapear os padrões de P bits é $P(P-1)+3$. Por exemplo, o mapeamento uniforme produz 59 rótulos de saída para as vizinhanças com oito pontos de amostragem e 243 rótulos para vizinhanças de 16 pontos de amostragem. Temos duas razões para omitir os padrões não-uniformes. A primeira é que a grande maioria dos padrões em imagens naturais são uniformes. OJALA; PIETIKAINEN; MAENPAA (2002) notaram em seus experimentos com imagens de texturas, padrões uniformes representam cerca de 90% de todos os padrões quando utilizamos a vizinhança (8,1) e cerca de 70% na vizinhança (16,2).

A Figura 3.6 mostra imagens retiradas da nossa base de dados divididas em imagens

destacando apenas os pixels com padrões uniformes (os pixels com padrões não-uniformes são preenchidos com a a cor branca) e imagens destacando somente pixels de padrões não-uniformes (os pixels com padrões uniformes são preenchidos com a a cor branca). Essas imagens são criadas utilizando o $LBP^{u2}_{8,1}$. A imagem com somente padrões uniformes contém uma quantidade consideravelmente maior de pixels. Outro fato impactante é que considerando apenas os pixels com padrões uniformes o plano de fundo é preservado. Isso acontece porque os pixels do plano de fundo têm todos a mesma cor (mesmo nível de cinza) e logo os padrões contêm pouca ou nenhuma transição. A segunda razão para considerar a utilização de padrões uniformes é a robustez estatística. A utilização de padrões uniformes ao invés de todos os padrões possíveis produziu melhores resultados de reconhecimento em diversas aplicações. Há indicações de que padrões uniformes são mais estáveis, i.e. menos suscetíveis a ruído, e além disso, considerar apenas padrões uniformes torna o número de rótulos LBP possíveis significativamente menor, fazendo com que uma estimativa confiável de suas distribuições requeira menos amostras. Os padrões uniformes permitem enxergar o método LBP como uma abordagem unificadora para os tradicionais e divergentes modelos estatístico e estrutural para análise de texturas (MAENPAA; PIETIKAINEN, 2005). Cada pixel é rotulado com o código da textura primitiva que mais se assemelha com a vizinhança local. Dessa forma, cada código LBP pode ser considerado como um *micro-texton*. Primitivos locais detectados pelo LBP incluem pontos, áreas planas, arestas, fim de arestas, curvas, etc. A combinação das abordagens estatística e estrutural apoia-se no fato de que a distribuição de micro-textons pode ser vista como regras de posicionamento estatístico. A distribuição LBP portanto, tem ambas as propriedades do método de análise estrutural: primitivos de textura e regras de posicionamento. Por outro lado, a distribuição é apenas uma estatística de imagem não linearmente filtrada, claramente tornando o método um método estatístico. Por essas razões, a distribuição LBP pode ser utilizada com sucesso para reconhecer uma vasta variedade de diferentes texturas, para as quais métodos estatísticos e estruturais normalmente são aplicados separadamente.

3.2.4 LBP Invariante a Rotação

Seja $U_p(n,r)$ uma representação que denota um padrão uniforme LBP específico. O par (n,r) específica um padrão uniforme de forma que n é número de bits "1"no padrão (corresponde ao número da linha na Figura 3.7) e r é a rotação do padrão (número da coluna na Figura 3.7) Agora se a vizinhança tem P pontos de amostragem, n recebe valores de 0 até $P+1$, onde $n = P+1$ é o rótulo especial que marca todos os padrões não-uniformes. Além disso, quando $1 \leq n \leq P-1$, a rotação do padrão está no intervalo $0 \leq r \leq P-1$. $I^{\alpha^\circ}(x,y)$ denota a rotação de imagem $I(x,y)$ por α graus. Sob essa rotação, o ponto (x,y) é rotacionado para a localização (x',y'). Uma vizinhança circular de amostras nos pontos $I(x,y)$ e $I^{\alpha^\circ}(x',y')$ também rotaciona por α°. Se as rotações são limitadas para múltiplos inteiros do ângulo entre dois pontos de amostra, i.e. $\alpha = \alpha\frac{360^\circ}{P}, \alpha = 0,1,\ldots,P-1$, isso rotaciona a vizinhança de amostragem por exatamente

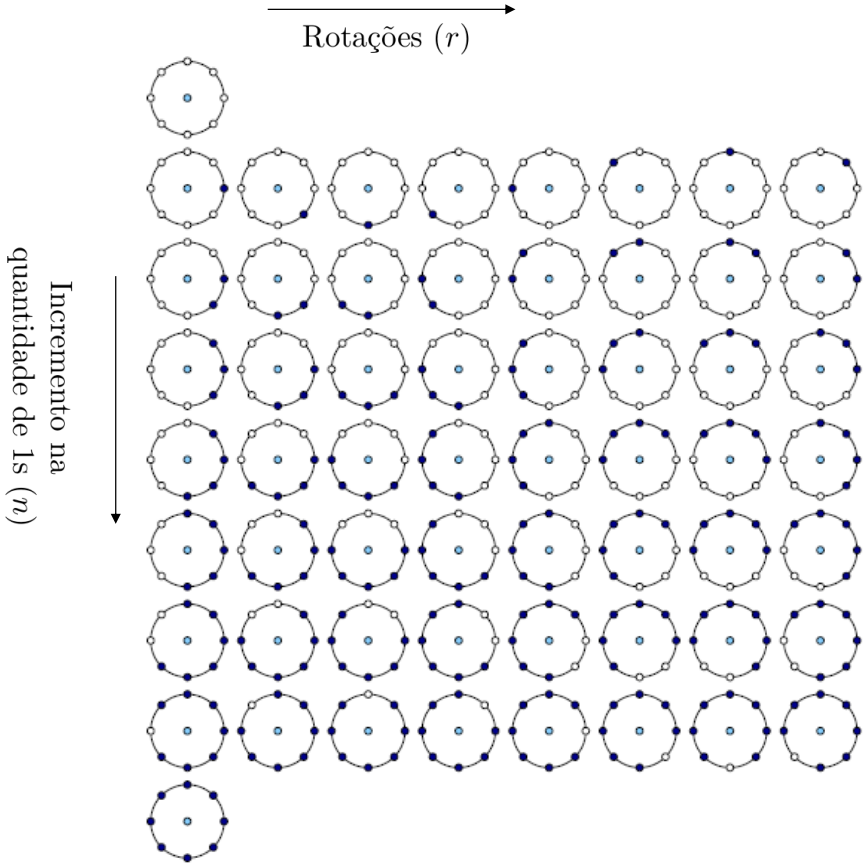


Figura 3.7: Os 58 padrões uniformes diferentes em uma vizinhança $(8,R)$. Adaptado de PIETIKAINEN et al. (2011)

α passos discretos.No entanto o padrão uniforme $U_P(n,r)$ no ponto (x,y) é substituído pelos padrões uniformes $U_P(n,r+amodP)$ no ponto (x',y') da imagem rotacionada. A partir dessa observação o LBP invariante a rotação foi introduzido em OJALA; PIETIKAINEN; MAENPAA (2002).

Rotações de uma imagem texturizada fazem com que os padrões LBP sejam transladados para uma localização diferente e rotacionados em relação a sua origem. Computando o histograma de códigos LBP o problema de transladação é normalizado. A normalização para a rotação é alcançada por um mapeamento invariante a rotação. Neste mapeamento, cada código binário LBP é rotacionado de forma circular par obter seu valor mínimo

$$LBP^{ri}_{P,R} = \min_i ROR(LBP_{P,R,i}) \tag{3.16}$$

onde $ROR(x,i)$ representa a rotação circular bit a bit para a direita da sequências de bits x com i passos. Para exemplificar, os códigos LBP de 8 bits 10000010b, 00101000b e 00000101b são todo mapeados para o código mínimo 00000101b. Omitindo artefatos de amostra, o histograma dos códigos $LBP^{ri}_{P,R}$ é invariante apenas para rotações de uma imagem por ângulos de $\alpha = \alpha\frac{360^\circ}{P}, \alpha = 0,1,\ldots,P-1$. No entanto experimentos de classificação mostram que este

descritor é muito robusto para rotações no mesmo plano de imagens em qualquer ângulo.

3.2.5 Variações do LBP e Aplicações

O sucesso dos métodos LBP em vários problemas de visão computacional e aplicações inspiraram uma grande quantidade de novas pesquisas em diferentes variações. Devido a sua flexibilidade o método LBP pode ser facilmente modificado para torná-lo adequado as necessidade de diferentes tipos de problemas. O LBP básico tem também alguns problemas que precisam ser levantados. Sendo assim, muitas extensões e modificações do LBP foram propostas como o objetivo de aumentar a robustez e poder discriminativo. Essas aplicações incluem principalmente reconhecimento de face (ZHANG; GAO, 2009), classificação de textura (LIAO; LAW; CHUNG, 2009), análise de imagens médicas (NANNI; LUMINI; BRAHNAM, 2010), busca de imagens (G. et al., 2010), reconhecimento de objetos (ZHU; BICHOT; CHEN, 2010) entre outras. Aparentemente o LBP ainda não foi utilizado para o problema de classificação de vídeos.

A escolha de um método apropriado para uma dada aplicação depende de muitos fatores, tais como poder de discriminação, eficiência computacional, robustez a variações de iluminação, e o sistema de obtenção de imagens utilizado. PIETIKAINEN et al. (2011) sugere que o LBP invariante a rotação oferece um bom ponto de partida quando se procura uma variante ótima para uma aplicação específica.

3.3 Combinação de Características

Fusão ou combinação de características é um método para combinar diferentes características extraídas de uma instância em uma base de dados, como por exemplo, cor e textura, no processo de classificação de imagens de esportes.

No nosso trabalho consideramos dois tipos de esquemas de combinação de características: *Early Fusion* e *Late Fusion*. *Early Fusion* realiza a combinação no espaço de características e possui um desempenho melhor para categorias que exibem constância em ambas cor e forma, enquanto que *late Fusion* realiza essa tarefa no espaço semântico e possui uma performance melhor quando uma das características se mantém constante e a outra varia de forma significativa (SNOEK; WORRING; SMEULDERS, 2005). Detalhes da nossa implementação são fornecidos no próximo capítulo.

3.3.1 *Early Fusion*

O método de (Early Fusion) é sensível em relação a dimensionalidade dos vetores de características. E apresenta como vantagem possuir apenas uma etapa de treinamento do modelo de aprendizagem. A combinação de característica é realizada concatenando o vetor de características normalizados representado pelo autocorrelograma e o vetor de características

representado pelo histograma de saída do LBP. Uma vez que obtemos essa nova representação multimodal das características, um classificador SVM é treinado para cada classe com esse novo conjunto de dados. A Figura 3.8 mostra uma visão geral do nosso esquema de *early fusion*

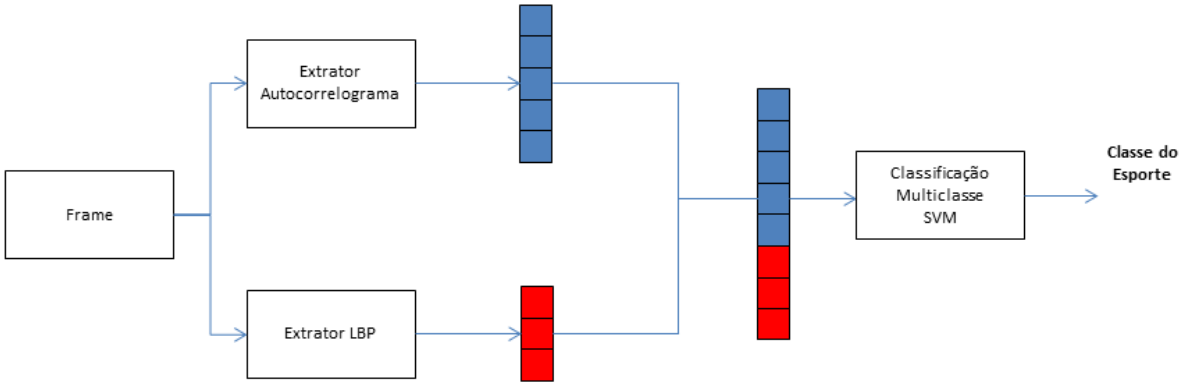


Figura 3.8: Visão geral da abordagem *early fusion*.

3.3.2 *Late Fusion*

O método de *late fusion* também é iniciado com a extração unimodal de características. Contrastando com o *early fusion* as predições dos classificadores são combinadas para determinar um score final de saída. Foi apresentada utilizando diferentes formalismos, tais como meta-classificação que busca reclassificar as classificações produzidas por outros classificadores (LIN; JIN; HAUPTMANN, 2002).

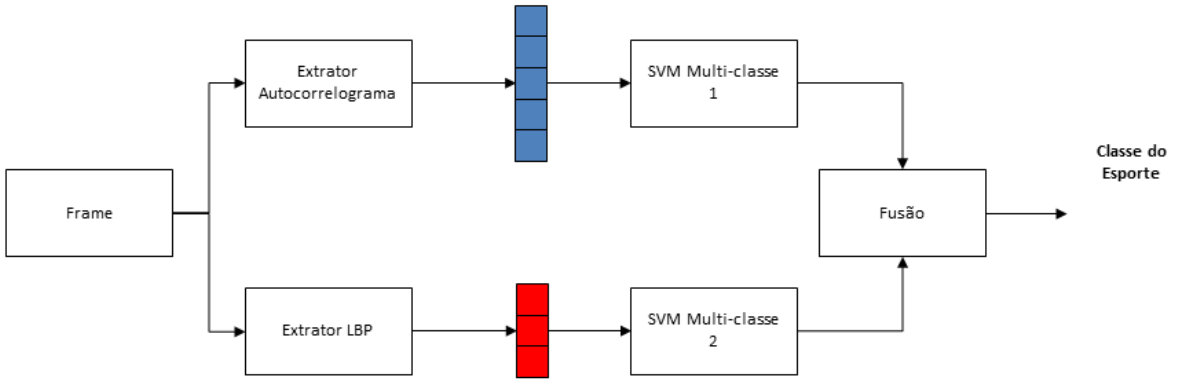


Figura 3.9: Visão geral da abordagem *late fusion*.

A teoria mais próxima de ilustrar uma *late fusion* são os métodos de empilhamento (*stacking*) que fazem parte dos métodos de comitês de classificadores ou *ensemble* (DIETTERICH, 2000). A ideia por trás de métodos de aprendizado *ensemble* (e.g. *bagging*, *boosting*, *stacking*) é melhorar a generalização treinando mais de um modelo em cada problema e combinar suas predições realizando a média, por votação ou por qualquer outro método (KITTLER et al., 1998). Uma grande desvantagem dos esquemas de *late fusion* é o seu custo em termos de esforço de

treinamento, pois cada característica requer uma etapa de aprendizado separada. Outra desvantagem dessa abordagem é potencial perda de correlação no espaço de características misturados, uma vez que os classificadores que estão mais alto na hierarquia trabalham com características de dimensão reduzida. Um esquema geral para *late fusion* é ilustrado na Figura 3.9.

4

Experimentos e Resultados

Este capítulo descreve os experimentos realizados para comparar o desempenho de diferentes abordagens para a classificação de vídeos de esportes: a utilização de apenas uma das características descritas no capítulo anterior por vez, e utilizando a abordagem de combinação de características. Dois tipos de combinação são utilizadas (*early fusion* e *late fusion*) para a avaliação de resultados.

Na seção 4.1 detalhamos a construção das base de vídeos contendo produções profissionais e gravações amadoras de esportes e justificamos a necessidade de criação de uma base própria ao invés da utilização de alguma base de vídeos da revisão de literatura.

Na seção 4.2 o processo de escolha de parâmetros é descrito para a técnica de classificação utilizando apenas autocorrelogramas, apenas o LBP e a combinação de características para treinamento e teste.

Na seção 4.3 a configuração do classificador SVM multi-classe é descrita e são fornecidos detalhes de sua implementação.

Na seção 4.4 descrevemos os experimentos realizados para avaliar o desempenho e construção da nossa abordagem.

Na última seção , é apresentada uma discussão a cerca dos resultados obtidos abordando as vantagens da técnica utilizada e a comparação com resultados de pesquisas semelhantes do estado da arte.

4.1 Base de Vídeos Amadores e Profissionais

Uma vez que não há uma base de vídeos pública de sequências de esportes gravadas por amadores, criamos nossa própria base de vídeos de esportes. A revisão dos estado da arte indica que todos os trabalhos produzem sua própria base de vídeos, e que muito provavelmente por questões legais não as disponibilizam, descrevendo apenas a natureza de produção das sequências (TV) e o meio de obtenção das mesmas (Internet). Avaliamos dois aspectos do processo de classificação. Primeiro classificando apenas os vídeos de esportes transmitidos pela televisão para analisar a performance e precisão do classificador. Em seguida realizamos uma segunda

análise utilizando vídeos amadores e profissionais de forma misturada para testar como isso afeta a taxa de classificação. Os experimentos foram realizados utilizando *10-fold cross validation* para o processo de classificação.

Neste trabalho dois tipos de bases de vídeos foram utilizados como mencionado anteriormente e descrito nas próximas subseções.

4.1.1 Base de Vídeos de TV

Uma base de vídeos balanceada (proporção quase idêntica entre as classes) com sequências de futebol, vôlei e tênis foram utilizados para os experimentos. A maior parte dos trabalhos analisados utiliza diferentes gêneros esportivos que podem ser classificados utilizando simples características de cor, tais como mergulho, tênis com quadra de saibro ou hóquei no gelo. Os vídeos de tênis utilizados em nossa base foram obtidos de jogos realizados em quadras de grama aumentando assim a dificuldade de discriminá-las das sequências de futebol mostrando o campo aberto. No total 20 horas de vídeos de esportes foram utilizadas das quais 70000 *frames* foram extraídos de maneira randômica. As 20 horas de vídeos foram compostas de 6 videoclipes de futebol, 6 videoclipes de voleibol e 6 videoclipes de tênis.

4.1.2 Base de Vídeos Gravados por Usuários

A base de vídeo amadores é composta por vídeos gravados por usuários de smartphones em partidas de futebol realizadas no mesmo estádio (*Arena das Dunas*, Natal-RN). Nenhuma orientação específica foi dada para os usuários amadores que foram instruídos apenas a gravar as partidas. Smartphones de diferentes fabricantes foram utilizados para coletar os vídeos de diferentes pontos de visualização marcados na Figura 4.1. O trabalho CRICRI et al. (2014) avalia a utilização de vídeos de diferentes ângulos na classificação de forma positiva, contribuindo para melhoras na taxa de acerto. No total, 20000 frames foram extraídos de 6 sequências de vídeos diferentes para a classificação.

4.2 Parâmetros e Configurações Para a Combinação de Características

Para realizar a classificação dos frames extraídos é necessário configurar alguns parâmetros das características utilizadas.

4.2.1 Parâmetro K do Autocorrelograma

Para diminuir a complexidade computacional do problema devido as dimensões das imagens, as sequências de vídeos são processadas para extrair frames redimensionados, diminuindo

Figura 4.1: Esquema gráfico da *Arena das Dunas* mostrando os ponto de visualização onde os vídeos foram gravados.

a resolução da imagem para 112x96 pixels e em seguida quantizamos a imagem para apenas 64 cores no conjunto de distâncias, i.e., $k \in \{1,3,5,7,9\}$. Cada valor de k gera um autocorrelograma para um frame que irá representar um vetor de características com 64 valores. Então os vetores de características para cada valor de k são concatenados para formar um vetor de característica maior com 320 (64 x 5 = 320) características. A escolha dos valores de k em WATCHARAPINCHAI et al. (2007) é feita de forma aparentemente arbitrária, mas uma análise exploratória mostrada para os nossos experimentos de classificação mostrou que a utilização combinada de todos os vetores de características concatenados para os valores de k provê resultados melhores do que a utilização individual dos mesmos. A Figura 4.2 descreve os resultado de uma validação da utilização dos vetores de características para os diferentes valores de k de forma separada ou concatenados, realizando a classificação dos autocorrelogramas na base de vídeos contendo sequências profissionais e amadoras.

Consideramos também na nossa análise a utilização de técnicas de seleção e redução de atributos para o pré-processamento dos autocorrelogramas utilizadas em XU et al. (2008), como forma de diminuir a complexidade de classificação. Conforme os experimentos realizados em WATCHARAPINCHAI et al. (2007), a utilização do PCA para redução de características reduziu a taxa de acerto da classificação e optamos por não considerá-lo em nossos experimentos.

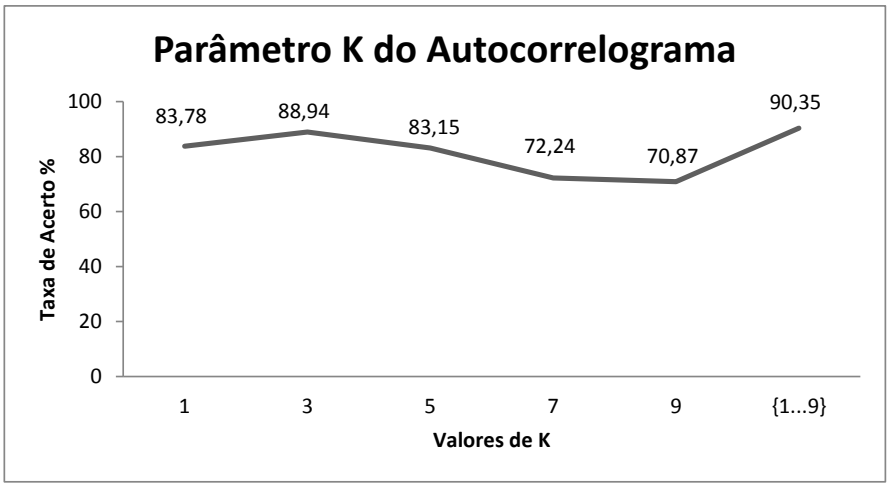


Figura 4.2: Resultado de classificação para vetores de características para cada valor do parâmetro k. A última entrada do gráfico, 1...9, representa a taxa de classificação para o vetor proveniente da concatenação dos demais.

4.2.2 Redução do Espaço de Cores para Autocorrelogramas

A quantização de cores é um dos recursos utilizados para reduzir o número de cores efetivamente utilizados no espaço de cores RGB e define os intervalos de cores que são representados por apenas uma cor escolhida de acordo com um critério específico. Para os autocorrelogramas, utilizamos uma quantização uniforme no cubo RGB, dividindo cada eixo em 4 partes e mapeando todas as cores dentro de um dos 64 sub-cubos formados nesse processo para a cor que está no centro deste sub-cubo. Tal processamento é chamado de **quantização uniforme** e permite que em todos os frames avaliados utilizemos as mesmas cores após a redução.

4.2.3 Configuração do LBP

Para a utilização do LBP é necessário primeiramente transformar os frames de entrada para imagens em nível de cinza. A conversão é realizada com uma transformação do espaço de cores *RGB* para o *HSV*, onde o canal de luminância V é considerado a imagem de saída em nível de cinza.

Nossa implementação do LBP como mencionado no capítulo 3 é representado por $LBP^{riu}_{8,1}$ significando que é invariante a rotação e utiliza uma vizinhança de raio 1 com 8 pontos de amostragem, considerando o mapeamento de padrões não-uniformes para um mesmo rótulo.

O LBP é extraído do mesmo frame que é redimensionado para 112x96 e utilizado para a extração do autocorrelograma, diminuindo também a complexidade computacional. O

histograma normalizado da imagem com os código LBP é o nosso vetor de saída com 59 características (58 provenientes de padrões uniformes e 1 para representar todos os padrões não-uniformes).

4.3 Classificação com SVM

Para o processo de classificação multi-classe, utilizamos *Support Vector Machines* (SVM), um reconhecido modelo de aprendizagem supervisionada para classificação binária que busca maximizar a margem de separação entre as instâncias de treinamento e a fronteira de decisão (BOSER; GUYON; VAPNIK, 1992). Dado um conjunto de treinamento com vetores de características (x_i, y_i) $i = 1, \ldots l$, onde $x_i \in R^n$ e $y \in \{0,1\}^l$, o SVM requer a solução do problema de otimização apresentado na equação 4.1:

$$\begin{aligned} \underset{w,b,\xi}{\text{minimiza}} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\ \text{sujeito a} \quad & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{4.1}$$

Dessa forma vetores de treinamento x_i são mapeados para um espaço de dimensionalidade maior (alguma vezes a dimensionalidade é infinita) pela função ϕ. O SVM encontra um hiperplano linear separatório com margem máxima nesse espaço de dimensionalidade maior. $C > O$ é o parâmetro de custo que regula o *trade-off* entre generalização e minimização de erro. Além disso, $K(x_i, x_j) = \phi(x_i^T)\phi(x_j)$ é chamada função de kernel. Existem 4 tipos básicos de funções de kernel:

- linear: $K(x_i, x_j) = x_i^T x_j$
- polinomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0.$
- *radial basis function* (RBF): $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0.$
- sigmoide: $K(x_i, x_j) = tanh(\gamma x_i^T x_j + r).$

Onde, γ, r, e d são parâmetros do kernel.

Para os nossos experimentos escolhemos SVMs com kernel RBF que segundo HSU; CHANG; LIN (2010), são indicados para problemas em que a quantidade de instâncias disponíveis para treinamento é muito maior que o número de características (no nosso caso da ordem de 100 vezes). Além disso uma grande quantidades de pesquisa na área de classificação de vídeos de esportes optam pela utilização de kernel RBF (WATCHARAPINCHAI et al., 2007; MOHAN; YEGNANARAYANA, 2010; WANG et al., 2004).

Para escolha de parâmetros do kernel utilizamos uma abordagem de *grid-search* para escolher os parâmetros C e γ. Os valores de C variaram em $\{1, 10, 100\}$ enquanto variamos o

valor de γ em $\{0.01, 0.1, 1, 10\}$. Utilizando validação cruzada com uma estratégia de $10-fold$ os valores escolhidos com a melhor taxa de classificação foram $C = 10$ e $\gamma = 0.1$.

4.3.1 Classificação Multiclasse com SVM

Para realizar a classificação multi-classe com SVM utilizamos a abordagem de um-contra-todos. A estratégia é treinar um modelo SVM para cada classe presente na base de treinamento (3 esportes + 1 uma classe de imagens de não-jogo), na tentativa de separar uma classe das restantes. Dessa forma passamos a ter 4 classificadores binários.

Cada classificador binário é construído com as seguintes especificações:

- A função de kernel escolhida para é a RBF (*Gaussian Radial Basis*).
- O parâmetro de custo $C = 10$ e $\gamma = 0.1$.
- O classificador fornece como saída uma estimativa de probabilidade e não uma classe.

Utilizamos o software LIBSVM (CHANG; LIN (2011)) para realizar a classificação fornecendo a saída como uma estimativa de probabilidade para predizer a classe das instâncias de teste. A classe assinalada é a do modelo que fornecer a maior probabilidade de predição.

4.3.2 Fusão de Classificadores

Em nossa implementação para realizar a abordagem de *late fusion* são treinados para cada conjunto de características $n(n-1)/2$ classificadores, onde n é o número de classes. Logo para nossa análise envolvendo 4 classes foram construídos 6 classificadores binários para cada conjunto de características (no nosso problemas são dois conjuntos: autocorrelogramas e LBP), totalizando doze classificadores.

Os classificadores são treinados para separar dados de cada classe a fim de que produzam um rótulo para uma instância de teste de uma classe ou da outra. A fusão é feita utilizando o voto da maioria e em caso de empate a saída dada pelos classificadores SVM treinados com os autocorrelogramas são fornecidos como saída porque apresentaram um resultado ligeiramente melhor nos experimentos, considerando as características de forma individual.

4.4 Avaliação da Classificação

Realizamos seis tipos de experimentos e avaliamos seus resultados para justificar a combinação de características realizadas. Cada experimento foi realizado 10 vezes com *seeds* diferentes e a métrica utilizada para avaliação de desempenho é a taxa de classificação, mas dispomos as matrizes de confusão para realizar uma análise mais detalhada. Os valores da taxa de classificação média são destacados em negrito juntamente com o desvio padrão e os resultados são sumarizados na Figura 4.4.

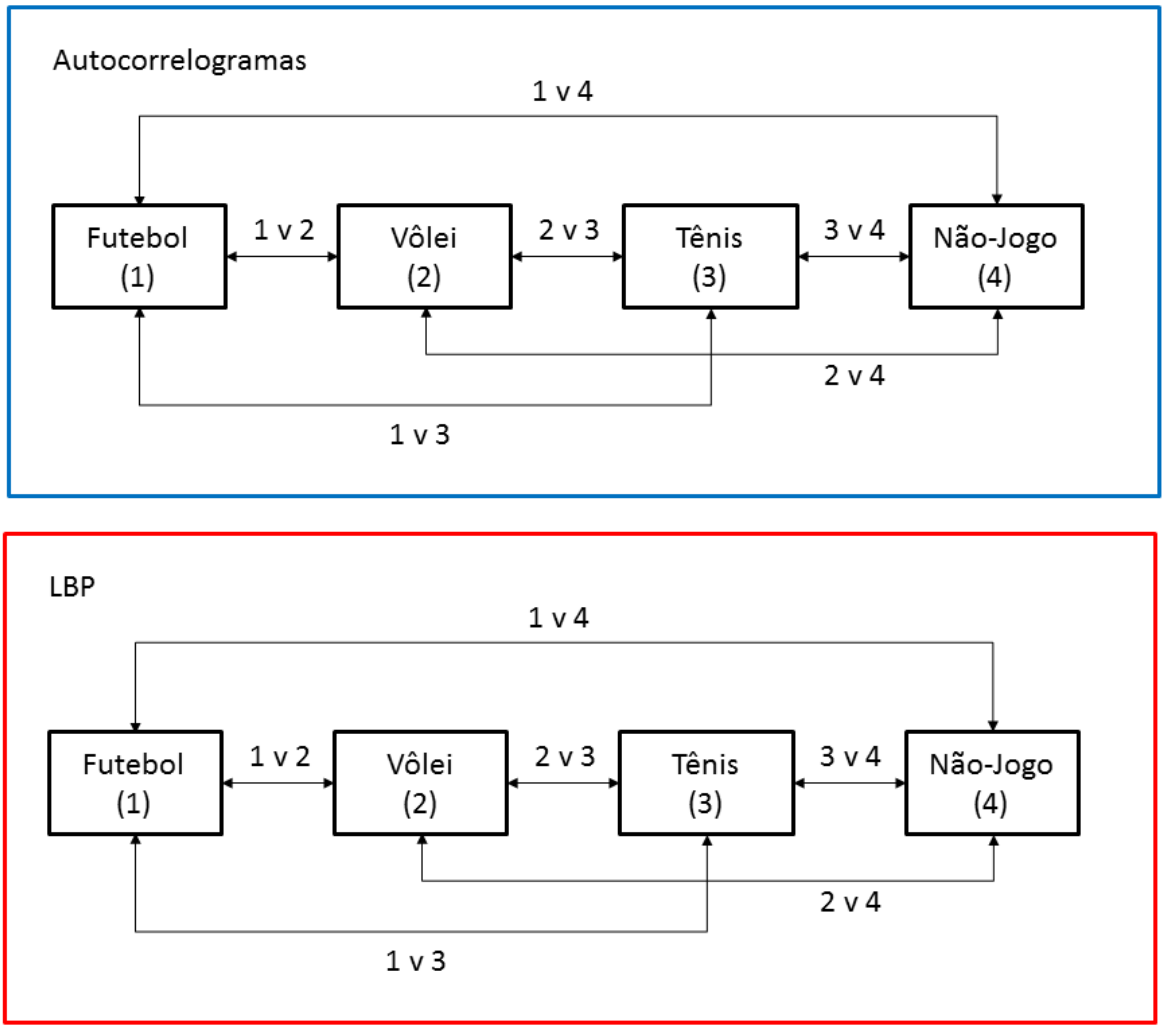


Figura 4.3: Ilustração do treinamento de classificadores SVM binários para posterior classificação com base no voto da maioria. Cada conjunto de características produz $n(n-1)/2$ classificadores, onde n é o número de classes.

4.4.1 Experimento 1 - Autocorrelogramas - 3 classes - TV

Nosso primeiro experimento realizado busca validar nossa implementação do extrator de autocorrelogramas implementado. Utilizamos para a classificação o SVM multiclasse com o intuito de classificar 3 tipos de esportes (Futebol, Voleibol e Tênis) mencionado na seção 4.3. A base de vídeos utilizada para esse experimento contém apenas vídeos provenientes de transmissões de televisão. A Tabela 4.1 mostra a matriz de confusão para esse primeiro experimento. A utilização dos autocorrelogramas como única características extraída apresenta bons resultados na classificação de vídeos de esportes como mostrado em WATCHARAPINCHAI et al. (2007). A taxa dos frames corretamente classificados é de **90,58** $\pm$ **0,71%**.

Tabela 4.1: Resultados da classificação da característica autocorrelograma utilizando a base de vídeos com imagens de TV e considerando três esportes.

Esportes	Futebol	Voleibol	Tênis
Futebol	90,93 ± 1,47 %	6,62 ± 1,41 %	2,45 ± 0,43 %
Voleibol	9,4 ± 0,75 %	87,3 ± 1,56 %	3,3 ± 1,74 %
Tênis	4,3 ± 2,45 %	2,6 ± 1,64 %	93,1 ± 1,64 %

4.4.2 Experimento 2 - Autocorrelogramas - 3 classes - TV + Amador

Para o segundo experimento realizamos novamente a classificação utilizando apenas autocorrelogramas e a mesma implementação do SVM multi-classe. No entanto, para esse experimento utilizamos a base de vídeos com esportes gravados de forma profissional (TV) e amadora (usuários de smartphones). A Tabela 4.2 mostra que não há uma mudança considerável quando utilizamos autocorrelogramas extraídos de vídeos de naturezas tão diferentes. A taxa de *frames* corretamente classificados é **90,35 ± 1,79 %**.

Tabela 4.2: Resultados da classificação da característica autocorrelograma utilizando a base de vídeos com imagens de TV e gravadas de smartphones, considerando três esportes.

Esportes	Futebol	Voleibol	Tênis
Futebol	90,58 ± 1,24 %	6,72 ± 0,88 %	2,7 ± 0,41 %
Voleibol	9,27 ± 0,88 %	87,23 ± 1,24 %	3,5 ± 0,47 %
Tênis	4,62 ± 1,11 %	2,47 ± 1,03 %	92,91 ± 0,38 %

4.4.3 Experimento 3 - Autocorrelogramas - 4 classes - TV + Amador

Para esse experimento passamos a considerar uma quarta classe para treinamento e teste do SVM. A quarta classe representa frames de não-jogo que em sua grande maioria são *close-ups* no rosto de atletas, frames que mostram os espectadores, vinhetas do patrocinador mostrando logomarcas, etc. A Figura 4.5 exemplifica *frames* de não jogo. A ideia é que realizando o treinamento de mais um modelo SVM para separar essas imagens tenhamos uma taxa de classificação global maior. A Tabela 4.3 mostra a matriz de confusão para a classificação. A taxa de *frames* corretamente classificados é **85,9 ± 2,49 %**.

Tabela 4.3: Resultados da classificação da característica autocorrelograma utilizando uma quarta classe para agrupar imagens de não-jogo. A base de vídeos utilizada é a de TV + imagens amadoras.

Esportes	Futebol	Voleibol	Tênis	Não Jogo
Futebol	88,52 ± 1,69 %	7,44 ± 1,01 %	2,36 ± 1,24 %	1,68 ± 0,65 %
Voleibol	15,33 ± 2,16 %	81,21 ± 0,64 %	2,76 ± 0,28 %	0,7 ± 0,15 %
Tênis	6,9 ± 0,39 %	2,3 ± 1,1 %	90,3 ± 0,35 %	0,5 ± 0,04 %
Não Jogo	17,7 ± 3,68 %	3,1 ± 1,15 %	5 ± 0,85 %	74,16 ± 1,92 %

4.4.4 Experimento 4 - LBP - 4 classes - TV + Amador

Replicamos o experimento 3 utilizando o extrator de características LBP. A Tabela 4.4 mostra que o LBP também é uma boa característica para classificação e aparentemente consegue representar bem as texturas dos campos e cena de jogo para os esportes, discriminando-os entre si e separando-os dos *frames* de não-jogo. No entanto, o contrário não acontece porque as cenas de não-jogo possuem grande variabilidade de cores mas pouca variação de textura (rostos e multidão), fazendo com que grande parte dos *frames* de não-jogo sejam classificados como esportes. A taxa de *frames* corretamente classificados para esse experimento é de **87,01 ± 0,6 %**.

Tabela 4.4: Resultados da classificação da característica *LBP* utilizando uma quarta classe para agrupar imagens de não-jogo. A base de vídeos utilizada é a de TV + imagens amadoras.

Esportes	Futebol	Voleibol	Tênis	Não Jogo
Futebol	97,81 ± 0,89 %	0,55 ± 0,16 %	1,64 ± 0,73 %	0 ± 0 %
Voleibol	5,56 ± 0,23 %	87 ± 0,41 %	7,44 ± 0,43 %	0 ± 0 %
Tênis	6,9 ± 0,28 %	2,3 ± 0,17 %	90,8 ± 0,18 %	0 ± 0 %
Não Jogo	11,8 ± 3,98 %	17,4 ± 3,75 %	37,5 ± 3,14 %	33,3 ± 5,41 %

4.4.5 Experimento 5 - Autocorrelogramas + LBP (Eearly Fusion) - 4 classes - TV + Amador

Avaliamos nesse experimento a combinação de características através da técnica de *early fusion* apresentada no capítulo 3. Podemos perceber uma melhora significativa quando combinamos os autocorrelogramas com o histograma LBP normalizado a nível de característica como podemos observar na Tabela 4.5. A taxa de *frames* corretamente classificados é **98,53 ± 0,83 %**.

Tabela 4.5: Resultados da classificação da características utilizando técnica de *early fusion* (Autocorrelograma + *LBP*) e utilizando uma quarta classe para agrupar imagens de não-jogo. A base de vídeos utilizada é a de TV + imagens amadoras.

Esportes	Futebol	Voleibol	Tênis	Não Jogo
Futebol	99,1 ± 0,61 %	0,33 ± 0,18 %	0,14 ± 0,05 %	0,43 ± 0,32 %
Voleibol	1,3 ± 0,33 %	98,14 ± 0,44 %	0,34 ± 0,04 %	0,22 ± 0,26 %
Tênis	0,37 ± 0,15 %	0,76 ± 0,39 %	98,87 ± 0,81 %	0 ± 0 %
Não Jogo	2,42 ± 0,41 %	0,7 ± 0,42 %	0 ± 0 %	96,88 ± 0,82 %

4.4.6 Experimento 6 - Autocorrelogramas + LBP (Late Fusion) - 4 classes - TV + Amador

Por fim o último experimento busca realizar a classificação com a estratégia de *late fusion* descrita na seção 4.3 deste capítulo. Embora a estratégia de *late fusion* apresente uma melhora quando comparada com a utilização de apenas uma das características para classificação o resultado é ligeiramente inferior ao da estratégia de *early fusion* e adiciona mais complexidade ao modelo, como mostra a Tabela 4.6. A taxa de *frames* corretamente classificados é **95,8** $\pm$ **0,4 %**.

Tabela 4.6: Resultados da classificação da características utilizando técnica de *late fusion* (Autocorrelograma + *LBP*) e utilizando uma quarta classe para agrupar imagens de não-jogo. A base de vídeos utilizada é a de TV + imagens amadoras.

Esportes	Futebol	Voleibol	Tênis	Não Jogo
Futebol	99,63 $\pm$ 0,21 %	0,04 $\pm$ 0,02 %	0,16 $\pm$ 0,08 %	0,17 $\pm$ 0,12 %
Voleibol	7,24 $\pm$ 0,82 %	92,64 $\pm$ 0,94 %	0,1 $\pm$ 0,07 %	0,02 $\pm$ 0,01 %
Tênis	2,58 $\pm$ 0,19 %	2,48 $\pm$ 0,61 %	94,94 $\pm$ 0,92 %	0 $\pm$ 0 %
Não Jogo	6,58 $\pm$ 0,41 %	1,4 $\pm$ 0,83 %	0,05 $\pm$ 0,02 %	91,97 $\pm$ 1,42 %

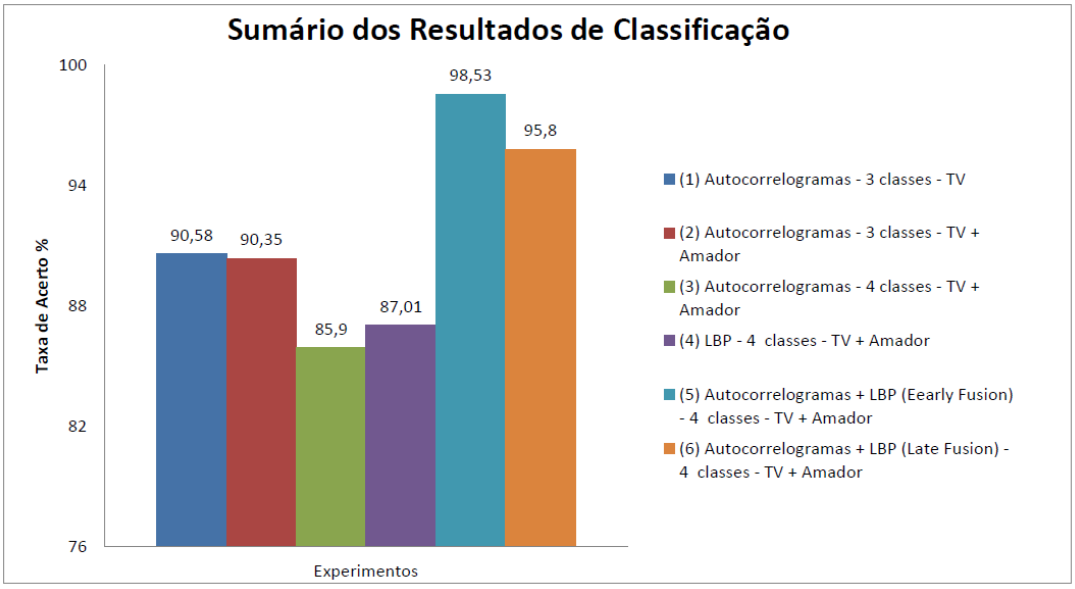


Figura 4.4: Resumo dos resultados de classificação dos experimentos realizados.

4.5 Discussão

Os resultados mostram que a abordagem de combinação dos autocorrelogramas com LBP proposta, realizando o *early fusion*, apresentam uma boa performance quando comparado com trabalhos do estado da arte em classificação de vídeos de esportes como mostrado na Tabela 4.7. Essa tabela é um complemento da Tabela 2.1 apresentando na terceira coluna o tamanhos

das bases de vídeos utilizadas e na quarta coluna a taxa de acerto da classificação. A diferença de representação na terceira coluna quanto ao tamanho da base vídeos é referente ao tipo de abordagens utilizadas na classificação. Pode-se classificar sequências de vídeos inteiros ou apenas *frames* extraídos dessas sequências. No nosso trabalho utilizamos aproximadamente 21 horas de vídeos das quais 90000 *frames* foram extraídos para realizar treinamento e teste. Alguns autores não informam o tamanho das bases utilizadas (N/A).

O resultado do método proposto não é o melhor para classificar a base de dados composta somente por vídeos transmitidos pela TV, mas é importante relembrar que nossa pesquisa realiza a classificação em uma base de vídeos profissionais e amadores. O único trabalho que realiza a classificação em vídeos gravados por usuários é apresentado em CRICRI et al. (2013) e obtém esses resultados utilizando uma combinação de características de vídeo, áudio e sensores de movimento auxiliares, enquanto que o trabalho proposto utiliza apenas características visuais. O trabalho apresentado em CRICRI et al. (2013) possui uma taxa de classificação de **81.82%** sem a utilização dos dados dos sensores auxiliares.

Devido à dificuldade de disponibilização das bases de treinamento e teste utilizadas na revisão de literatura, a comparação indireta é o indicativo mais viável do desempenho da técnica implementada. Comparações semelhantes são apresentadas em ZHANG et al. (2012), LI et al. (2009) e WANG et al. (2004). Nosso objetivo não é superar o desempenho de outras técnicas, mas mostrar as vantagens de nossa implementação quando vídeos de natureza amadora e profissional são classificados.

Os *frames* dos vídeos transmitidos pela TV mostrados na Figura 4.5 são difíceis de classificar utilizando apenas autocorrelogramas porque diferem muito das sequências de jogo corrido (sendo difícil até para um ser humano identificar um esporte através de imagens dos espectadores). *Close-ups* e sequências mostrando os espectadores não estão correlacionadas com as regras do jogo. No experimento 3 tentamos adicionar uma quarta classe para agrupar esses *frames* difíceis de classificar, mas os resultados não foram satisfatórios, provavelmente porque os *frames* de não-jogo dos três esportes utilizados apresentam uma variabilidade muito grande.

Por outro lado os autocorrelogramas mostram sua robustez em casos como os mostrados na Figura 4.6, com refletores de luz e oclusão parcial ocasionada por pessoas passando em frente a câmera, classificando corretamente esses *frames*. Até mesmo *frames* distorcidos por movimentos bruscos da câmera são corretamente classificados porque mantêm a correlação espacial entre as cores da imagem.

O LBP no experimento 4, utilizado sozinho possui uma boa taxa de classificação, mas tem ainda mais dificuldade de separar os *frames* de não-jogo. A combinação das características mostradas nos experimentos 5 e 6 mostraram o aumento significativo na taxa de classificação quando utilizamos um modelo simples de combinação de características.

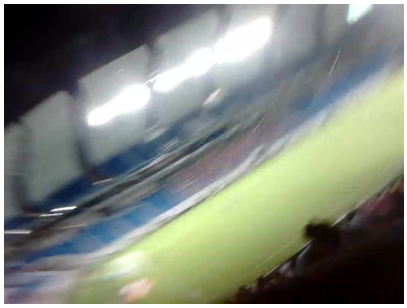
Tabela 4.7: Tamanho das bases de vídeos e resultados dos trabalhos relacionados.

Referência	Esportes	Tamanho	Resultado
YUAN et al. (2006)	6	60 horas	94.71%
DONG et al. (2012)	10	Aprox. 171 horas	87.3%
GIBERT; LI; DOERMANN (2003)	4	220 min.	93%
MOHAN; YEGNANARAYANA (2010)	5	5h. 30 min.	94.4%
LEE; HOFF (2007)	2	Aprox. 1 hora	94.2%
LI et al. (2009)	14	114 horas	88.8%
MUTCHIMA; SANGUANSAT (2012)	20	200 min.	96.65%
SIGARI; SURESHJANI; SOLTANIAN-ZADEH (2011)	7	(104 video clips)	78.8%
WANG et al. (2004)	4	(173 clipes de teste)	88%
WANG; XU; CHNG (2006)	3	16 horas	100%
WATCHARAPINCHAI et al. (2007)	7	233 min.	91.1%
XU et al. (2008)	4	1200 frames	N/A
YUAN; WAN (2004)	5	N/A	97.1%
CRICRI et al. (2013)	6	Aprox. 73 horas	90.91%
CRICRI et al. (2014)	6	Aprox. 73 horas	95.45%
Método Proposto	3	90000 frames	98.53%

Figura 4.5: *Close-ups*, espectadores, multidão e vinhetas do patrocinador: sequências de não-jogo produzem *frames* que dificultam a classificação.

(a) Oclusão parcial.

(b) Refletores na imagem.



(c) Distorção causada por movimentação brusca da câmera.

Figura 4.6: Frames corretamente classificados pela abordagem utilizando autocorrelogramas.

5

Conclusão

Este trabalho apresenta uma revisão dos principais estudos do estado da arte para a tarefa de classificação de vídeos de esportes com o intuito de propor uma técnica que possa lidar como a classificação simultânea de vídeos profissionais e amadores, uma vez que estes possuem diferentes características. Nesta revisão descrevemos as principais abordagens e as características visuais e acústicas já extraídas para a realização dessa tarefa, mas optamos por desenvolver uma técnica que não utilizasse áudio dos vídeos devido a grande variabilidade introduzida pelo ruído em gravações amadoras.

Nossa abordagem utiliza principalmente duas características:

- Uma característica visual de cor, os autocorrelogramas, que adicionam informação espacial a distribuição de cores, sendo portanto mais eficazes que histogramas utilizados em outras abordagens. Autocorrelogramas já foram utilizados em outros trabalho para o problema de classificação de vídeos de esportes em bases de vídeos profissionais e obtiveram bons resultados na classificação através da extração de *frames*.

- E uma característica de textura, os *local binary patterns* (LBP) uma técnica que descreve a textura através de diferenças de contraste de padrões espaciais locais. O LBP é muito utilizado em diversos áreas da visão computacional, mas até o presente momento nossa pesquisa não encontrou trabalhos relacionados a sua utilização para classificação de vídeos de esportes.

Dessa forma propomos um método robusto ante as diferenças de produção entre vídeos amadores e profissionais para discriminar entre três esportes (futebol, vôlei e tênis). Uma quarta classe que engloba imagens que não representam cenas de jogo (*close-ups*, imagens dos espectadores, vinhetas do patrocinador) também é utilizada como forma de facilitar o processo de classificação. Nossa base de vídeo é composta por 6 sequências de vídeos profissionais de cada esporte num total de 70000 *frames* extraídos de vídeos profissionais e 6 sequências de vídeos amadores de futebol totalizando 20000 *frames* extraídos de vídeos amadores gravados na *Arena da Dunas (Natal-RN)*. Como métrica de avaliação dos experimentos utilizamos a taxa de

acerto de classificação de *frames* por ser a mesma métrica utilizadas nas pesquisas do estado da arte.

Antes da realização do experimentos buscamos ajustar parâmetros para ambas as técnicas. Para o autocorrelograma utilizamos diferentes valores para o parâmetro k para formar vetores de características de maior dimensão que são mais robustos a ruído, o que implica numa melhor taxa de classificação. Dentre as inúmeras versões do LBP utilizamos a que considera os padrões uniformes e é invariante a rotação por ser tratar de uma versão robusta e indicada pela revisão de literatura como ponto de partida na utilização do LBP.

O modelo de classificação utilizado foi o SVM com função de *kernel* RBF, um modelo bastante utilizado em classificação de vídeos de esportes na revisão de literatura. Os parâmetros para esse modelos foram escolhidos como sendo $C = 10$ e $\gamma = 0.1$ através de uma abordagem de *grid-search*. Alguns trabalhos avaliaram a utilização de técnicas de redução de dados como o PCA para pré-processamento dos vetores de características mas chegaram a conclusões que a redução impacta negativamente na taxa de acerto.

Os experimentos foram realizados em seis etapas descritas brevemente a seguir para validar a nossa abordagem:

- A classificação das 3 classes de esportes utilizando apenas os autocorrelogramas extraídos de vídeos profissionais, resultando em uma taxa de acerto de 90,58%. Esse resultado mostrou que os autocorrelogramas utilizados sozinhos já apresentam boas taxas de classificação quando vídeos profissionais são utilizados.

- A classificação das 3 classes de esportes utilizando apenas os autocorrelogramas extraídos de vídeos amadores e profissionais, resultando em uma taxa de acerto de 90,35%. Com essa taxa de classificação percebemos que a introdução dos vídeos amadores não afetou drasticamente os resultados obtidos.

- A classificação das 3 classes de esportes mais a classe de cenas de não-jogo utilizando apenas os autocorrelogramas extraídos de vídeos amadores e profissionais, resultando em uma taxa de acerto de 85,9%, diminuindo assim a taxa de classificação porque mais *frames* extraídos dos vídeos foram classificados como *frames* de não-jogo.

- A classificação das 3 classes de esportes mais a classe de cenas de não-jogo utilizando apenas os padrões LBP extraídos de vídeos amadores e profissionais, resultando em uma taxa de acerto de 87,01%. Esse experimento foi necessário para avaliar o desempenho do LBP utilizado sozinho na classificação e mostrar que o autocorrelograma e o LBP classificam os *frames* de maneira diferente.

- A classificação das 3 classes de esportes mais a classe de cenas de não-jogo utilizando autocorrelograma e LBP extraídos de vídeos amadores e profissionais utilizando a abordagem de *early fusion*, resultando em uma taxa de acerto de 98,53%. Notavelmente a taxa de acerto aumentou nesse experimento mostrando que a combinação de

cor e textura melhoram os resultados de classificação como uma combinação simples de junção de características.

- A classificação das 3 classes de esportes mais a classe de cenas de não-jogo utilizando Autocorrelograma e LBP extraídos de vídeos amadores e profissionais utilizando a abordagem de *late fusion*, resultando em uma taxa de acerto de 95,8%. A combinação *late fusion* também obteve bons resultados combinando a saída de classificação de vários classificadores SVM.

Os resultados obtidos mostraram que ambas as técnicas apresentam bons resultados quando realizam a classificação utilizando modelos SVM multi-classe em uma base de vídeos misturados, demonstrando sua robustez em relação a variação de movimento de câmera e cor de diferentes dispositivos em gravações amadoras e profissionais. Ocorre uma melhora significativa quando as característica são combinadas por métodos de *early fusion* e *late fusion*, tendo a primeira apresentado resultados melhores e menor complexidade que a segunda.

5.1 Limitações

Uma das grandes limitações do trabalho desenvolvido é a indisponibilidade de bases de vídeos das pesquisas do estado da arte para a comparação de desempenho da nossa abordagem. Esses vídeos são apenas descritos de maneira superficial como sendo vídeos de transmissões de esportes pela televisão e obtidos na internet. Essa limitação torna difícil realizar comparações com outras abordagens, mas nosso objetivo desde o início do trabalho era desenvolver uma maneira robusta de classificar esportes em uma base de vídeos misturada. Comparando nosso resultado com pesquisas do estado da arte que realizam a classificação de esportes em vídeos amadores, quando utilizam apenas características visuais e de áudio, a melhor taxa de acerto de outras técnicas é cerca de 81,82%. Dessa forma percebemos melhoras que sugerem que tais características podem se incorporadas em outras abordagens para obtenção de ganhos na taxa de acerto.

5.2 Contribuições e Trabalhos Futuros

O desenvolvimento de uma técnica robusta para classificação de vídeos de esportes amadores e profissionais é tido como a principal das contribuições. Essa abordagem aumenta o leque de aplicações que podem ser desenvolvidas em cima de informação misturadas sem conhecimento prévio de domínio. O trabalho também introduz a utilização do LBP como uma boa característica a ser utilizada em classificação de vídeos de esportes. Vários trabalhos da revisão de literatura utilizam características de cor que podem ser combinadas com o LBP para aumentar a taxa de acerto. Outra contribuição é a formação inicial de uma base de vídeos que pode ser disponibilizada publicamente e posteriormente estendida uma vez que nenhuma

pesquisa na área disponibiliza a base utilizada, descrevendo apenas a sua natureza e meio de obtenção.

Como trabalhos futuros, incluímos a adição de novos esportes com mais vídeos amadores à nossa base de dados. Uma base maior deve atrair pesquisadores para construção de técnicas e abordagens revelando novos aspectos a serem explorados na classificação de vídeos de esportes. Destacamos também como uma linha de pesquisa a ser avaliada, o estudo da realização da classificação em duas etapas distintas, primeiramente classificando os vídeos em amadores e profissionais através da extração de características adequadas, para só então classificar o gênero esportivo.

Ainda como trabalho futuro existem pontos que podem ser otimizados na técnica implementada. A revisão de literatura mostra que o cálculo de distância entre *pixels* utilizado nos autocorrelogramas considera apenas a distância L_∞-norm quando outros tipos de distâncias, como por exemplo a euclidiana, poderiam ser utilizadas para avaliar o impacto no desempenho da técnica. Além disso, existem diferentes versões e adaptações do LBP a fim de otimizar seu uso para um domínio específico. Novos experimentos podem ser realizados para avaliar o desempenho de alguma dessas variações. A combinação de características *late fusion* pode ser mais explorada com técnicas dos estado da arte como *random forests* e utilizando comitês de classificadores variados.

Referências

AMIRI, A.; FATHY, M. Video Shot Boundary Detection Using QR-Decomposition and Gaussian Transition Detection. **EURASIP Journal on Advances in Signal Processing**, [S.l.], v.2009, n.1, 2009.

BELHUMEUR, P. N.; HESPANHA, J. P.; KRIEGMAN, D. J. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Washington, DC, USA, v.19, n.7, p.711–720, July 1997.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A Training Algorithm for Optimal Margin Classifiers. In: FIFTH ANNUAL WORKSHOP ON COMPUTATIONAL LEARNING THEORY, New York, NY, USA. **Proceedings...** ACM, 1992. p.144–152. (COLT '92).

BREZEALE, D.; COOK, D. Automatic Video Classification: a survey of the literature. **Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on**, [S.l.], v.38, n.3, p.416–430, May 2008.

CANNY, J. A Computational Approach to Edge Detection. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, [S.l.], v.PAMI-8, n.6, p.679–698, Nov 1986.

CHANG, C.-C.; LIN, C.-J. LIBSVM: a library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**, [S.l.], v.2, p.27:1–27:27, 2011.

CRICRI, F. et al. Multi-sensor fusion for sport genre classification of user generated mobile videos. In: MULTIMEDIA AND EXPO (ICME), 2013 IEEE INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2013. p.1–6.

CRICRI, F. et al. Sport Type Classification of Mobile Videos. **Multimedia, IEEE Transactions on**, [S.l.], v.16, n.4, p.917–932, June 2014.

DIETTERICH, T. G. Ensemble Methods in Machine Learning. In: FIRST INTERNATIONAL WORKSHOP ON MULTIPLE CLASSIFIER SYSTEMS, London, UK. **Proceedings...** Springer-Verlag, 2000. p.1–15.

DONG, Y. et al. Automatic sports video genre categorization for broadcast videos. In: VISUAL COMMUNICATIONS AND IMAGE PROCESSING (VCIP), 2012 IEEE. **Anais...** [S.l.: s.n.], 2012. p.1–5.

G., C. et al. Large Scale Online Learning of Image Similarity Through Ranking. **Journal of Machine Learning Research**, [S.l.], v.11, p.1109–1135, 2010.

GADE, R.; MOESLUND, T. Sports Type Classification Using Signature Heatmaps. In: COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS (CVPRW), 2013 IEEE CONFERENCE ON. **Anais...** [S.l.: s.n.], 2013. p.999–1004.

GIBERT, X.; LI, H.; DOERMANN, D. Sports video classification using HMMS. In: MULTIMEDIA AND EXPO, 2003. ICME '03. PROCEEDINGS. 2003 INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2003. v.2, p.345–348.

GUO, J. et al. Detecting complex events in user-generated video using concept classifiers. In: CONTENT-BASED MULTIMEDIA INDEXING (CBMI), 2012 10TH INTERNATIONAL WORKSHOP ON. **Anais...** [S.l.: s.n.], 2012. p.1–6.

HARALICK, R. Statistical and structural approaches to texture. **Proceedings of the IEEE**, [S.l.], v.67, n.5, p.786–804, May 1979.

HESSELDAHL, A. **Cisco**: the internet is, like, really big, and getting bigger. [Online; Accessado em 2014-07-2], `http://allthingsd.com/20110601/cisco-the-internet-is-like-really-big-and-getting-bigger/`.

HSU, C. wei; CHANG, C. chung; LIN, C. jen. **A practical guide to support vector classification**. 2010.

HU, W. et al. A Survey on Visual Content-Based Video Indexing and Retrieval. **Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on**, [S.l.], v.41, n.6, p.797–819, Nov 2011.

HUANG, J. et al. Image indexing using color correlograms. In: COMPUTER VISION AND PATTERN RECOGNITION, 1997. PROCEEDINGS., 1997 IEEE COMPUTER SOCIETY CONFERENCE ON. **Anais...** [S.l.: s.n.], 1997. p.762–768.

KITTLER, J. et al. On combining classifiers. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, [S.l.], v.20, n.3, p.226–239, Mar 1998.

LEE, J. Y.; HOFF, W. Activity Identification Utilizing Data Mining Techniques. In: MOTION AND VIDEO COMPUTING, 2007. WMVC '07. IEEE WORKSHOP ON. **Anais...** [S.l.: s.n.], 2007. p.12–23.

LI, L. et al. Automatic sports genre categorization and view-type classification over large-scale dataset. In: ACM MULTIMEDIA. **Anais...** ACM, 2009. p.653–656.

LIAO, S.; LAW, M.; CHUNG, A. Dominant Local Binary Patterns for Texture Classification. **Image Processing, IEEE Transactions on**, [S.l.], v.18, n.5, p.1107–1118, May 2009.

LIN, W.-H.; JIN, R.; HAUPTMANN, A. G. Meta-Classification of Multimedia Classifiers. In: INTERNATIONAL WORKSHOP ON KNOWLEDGE DISCOVERY IN MULTIMEDIA AND COMPLEX DATA (KDMCD 2002, Taipei, Taiwan. **Anais...** [S.l.: s.n.], 2002.

MAENPAA, T.; PIETIKAINEN, M. Texture analysis with local binary patterns. In: CHEN, C. H.; PAU, L. F.; WANG, P. S. P. (Ed.). **Handbook of Pattern Recognition and Computer Vision**. 3.ed. Singapore: World Scientific Publishing Co., Inc., 2005. p.197–216.

MOHAN, C. K.; YEGNANARAYANA, B. Classification of sport videos using edge-based features and autoassociative neural network models. **Signal, Image and Video Processing**, [S.l.], v.4, n.1, p.61–73, 2010.

MUTCHIMA, P.; SANGUANSAT, P. TF-RNF: a novel term weighting scheme for sports video classification. In: SIGNAL PROCESSING, COMMUNICATION AND COMPUTING (ICSPCC), 2012 IEEE INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2012. p.244–249.

NANNI, L.; LUMINI, A.; BRAHNAM, S. Local binary patterns variants as texture descriptors for medical image analysis. **Artificial Intelligence in Medicine**, [S.l.], v.49, n.2, p.117 – 125, 2010.

NIU, Y.; LIU, F. What Makes a Professional Video? A Computational Aesthetics Approach. **Circuits and Systems for Video Technology, IEEE Transactions on**, [S.l.], v.22, n.7, p.1037–1049, July 2012.

OJALA, T. et al. Texture discrimination with multidimensional distributions of signed gray-level differences. **Pattern Recognition**, [S.l.], v.34, n.3, p.727 – 739, 2001.

OJALA, T.; PIETIKAINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. **Pattern Recognition**, [S.l.], v.29, n.1, p.51–59, Jan. 1996.

OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, [S.l.], v.24, n.7, p.971–987, Jul 2002.

PIETIKAINEN, M. et al. **Local Binary Patterns for Still Images**. [S.l.]: Springer, 2011. 13-47p. v.40.

SIGARI, M.; SURESHJANI, S.; SOLTANIAN-ZADEH, H. Sport Video Classification Using an Ensemble Classifier. In: MACHINE VISION AND IMAGE PROCESSING (MVIP), 2011 7TH IRANIAN. **Anais...** [S.l.: s.n.], 2011. p.1–4.

SKULSUJIRAPA, P.; ARAMVITH, S.; SIDDHICHAI, S. Development of digital image retrieval technique using autocorrelogram and wavelet based texture. In: CIRCUITS AND SYSTEMS, 2004. MWSCAS '04. THE 2004 47TH MIDWEST SYMPOSIUM ON. **Anais...** [S.l.: s.n.], 2004. v.1, p.273–276.

SNOEK, C. G. M.; WORRING, M.; SMEULDERS, A. W. M. Early Versus Late Fusion in Semantic Video Analysis. In: ANNUAL ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 13., New York, NY, USA. **Proceedings...** ACM, 2005. p.399–402.

SUGANO, M. et al. Genre classification method for home videos. In: MULTIMEDIA SIGNAL PROCESSING, 2009. MMSP '09. IEEE INTERNATIONAL WORKSHOP ON. **Anais...** [S.l.: s.n.], 2009. p.1–5.

TAKAGI, S. et al. Sports video categorizing method using camera motion parameters. In: MULTIMEDIA AND EXPO, 2003. ICME '03. PROCEEDINGS. 2003 INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2003. v.2, p.461–464.

TAN, Y.-P. et al. Rapid estimation of camera motion from compressed video with application to video annotation. **Circuits and Systems for Video Technology, IEEE Transactions on**, [S.l.], v.10, n.1, p.133–146, Feb 2000.

WANG, D.-H. et al. News sports video shot classification with sports play field and motion features. In: IMAGE PROCESSING, 2004. ICIP '04. 2004 INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2004. v.4, p.2247–2250 Vol. 4.

WANG, J.; XU, C.; CHNG, E. Automatic Sports Video Genre Classification using Pseudo-2D-HMM. In: PATTERN RECOGNITION, 2006. ICPR 2006. 18TH INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2006. v.4, p.778–781.

WATCHARAPINCHAI, N. et al. A discriminant approach to sports video classification. In: COMMUNICATIONS AND INFORMATION TECHNOLOGIES, 2007. ISCIT '07. INTERNATIONAL SYMPOSIUM ON. **Anais...** [S.l.: s.n.], 2007. p.557–561.

WU, P.-H. et al. Separation of Professional and Amateur Video in Large Video Collections. In: PACIFIC RIM CONFERENCE ON MULTIMEDIA: ADVANCES IN MULTIMEDIA INFORMATION PROCESSING, 10., Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2009. p.33–44.

XIONG, Z. Audio-visual Sports Highlights Extraction Using Coupled Hidden Markov Models. **Pattern Anal. Appl.**, London, UK, v.8, n.1, p.62–71, Sept. 2005.

XU, M. et al. Comparison analysis on supervised learning based solutions for sports video categorization. In: MULTIMEDIA SIGNAL PROCESSING, 2008 IEEE 10TH WORKSHOP ON. **Anais...** [S.l.: s.n.], 2008. p.526–529.

YOUTUBE. **Statistics**. [Online; Accessado em 2014-07-10], `https://www.youtube.com/yt/press/statistics.html`.

YUAN, X. et al. Automatic Video Genre Categorization using Hierarchical SVM. In: IMAGE PROCESSING, 2006 IEEE INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2006. p.2905–2908.

YUAN, Y.; WAN, C. The application of edge feature in automatic sports genre classification. In: CYBERNETICS AND INTELLIGENT SYSTEMS, 2004 IEEE CONFERENCE ON. **Anais...** [S.l.: s.n.], 2004. v.2, p.1133–1136.

ZHANG, N. et al. A Generic Approach for Systematic Analysis of Sports Videos. **ACM Transactions on Intelligent Systems and Technology**, New York, NY, USA, v.3, n.3, p.46:1–46:29, May 2012.

ZHANG, X.; GAO, Y. Face Recognition Across Pose: a review. **Pattern Recognition**, New York, NY, USA, v.42, n.11, p.2876–2896, Nov. 2009.

ZHANG, Y.; JIA, Z.; CHEN, T. Image retrieval with geometry-preserving visual phrases. In: COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2011 IEEE CONFERENCE ON. **Anais...** [S.l.: s.n.], 2011. p.809–816.

ZHU, C.; BICHOT, C.-E.; CHEN, L. Multi-scale Color Local Binary Patterns for Visual Object Classes Recognition. In: PATTERN RECOGNITION (ICPR), 2010 20TH INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2010. p.3065–3068.