SUMARIZAÇÃO AUTOMÁTICA EM MELHORES MOMENTOS DE TRANSMISSÕES TELEVISIVAS DE FUTEBOL

Luiz Gabriel Lins Bentes Mendonça de Vasconcelos

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientadores: Sergio Lima Netto
Eduardo Antônio Barros da Silva

Rio de Janeiro
Junho de 2011

SUMARIZAÇÃO AUTOMÁTICA EM MELHORES MOMENTOS DE TRANSMISSÕES TELEVISIVAS DE FUTEBOL

Luiz Gabriel Lins Bentes Mendonça de Vasconcelos

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Sergio Lima Netto, Ph.D.

Prof. Eduardo Antônio Barros da Silva, Ph.D.

Prof. José Gabriel Rodriguez Carneiro Gomes, Ph.D.

Prof. Junior Barrera, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2011

Vasconcelos, Luiz Gabriel Lins Bentes Mendonça de

Sumarização Automática em Melhores Momentos de Transmissões Televisivas de Futebol/Luiz Gabriel Lins Bentes Mendonça de Vasconcelos. – Rio de Janeiro: UFRJ/COPPE, 2011.

VIII, 87 p. 29, 7cm.

Orientadores: Sergio Lima Netto

Eduardo Antônio Barros da Silva

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2011.

Referências Bibliográficas: p. 78 – 87.

1. Cores Dominantes. 2. Correlação por Fase. 3. Pitch e Energia de Áudio. 4. AdaBoost. I. Netto, Sergio Lima *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Agradecimentos

Meus sinceros agradecimentos:

- aos professores Sergio Lima Netto e Eduardo Antônio Barros da Silva, pela orientação dada durante todo o período do mestrado, pelas oportunidades oferecidas, pela paciência e sabedoria em momentos importantes;
- aos professores José Gabriel Rodrigues Carneiro Gomes e Junior Barrera, por aceitarem o convite de participação na banca de examinação deste trabalho;
- à minha namorada, pelo seu amor, apoio e compreensão nos momentos de ausência;
- à minha família, pelo apoio incondicional;
- aos meus colegas de turma, pela amizade e suporte durante esses anos de mestrado;
- à Central Globo de Engenharia da TV Globo e seus funcionários, pelo material fornecido e toda ajuda necessária ao desenvolvimento deste projeto; e
- a todos funcionários e professores do Programa de Engenharia Elétrica da COPPE, por todos os serviços prestados.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

SUMARIZAÇÃO AUTOMÁTICA EM MELHORES MOMENTOS DE TRANSMISSÕES TELEVISIVAS DE FUTEBOL

Luiz Gabriel Lins Bentes Mendonça de Vasconcelos

Junho/2011

Orientadores: Sergio Lima Netto
Eduardo Antônio Barros da Silva

Programa: Engenharia Elétrica

As recentes evoluções na captura, armazenamento e recuperação de vídeo têm acarretado em um aumento substancial na quantidade de conteúdo multimídia relacionado a esportes. Além disso, provedores de conteúdo têm expandido a disponibilidade de eventos esportivos a serem transmitidos para todo mundo em resposta ao significativo aumento de interesse de audiência. Assim, essas tendências apontam para a necessidade de desenvolvimento de ferramentas efetivas e eficientes para identificar melhores momentos nesses eventos, e, desta forma, reduzir esforços de telespectadores e empresas ao buscar o que lhes interessa. Este trabalho apresenta o desenvolvimento de um algoritmo para identificação automática de melhores momentos em vídeos de partidas de futebol transmitidas pela TV, onde os melhores momentos são definidos pelos gols e as ameaças de gol que ocorreram na partida. No algoritmo proposto, características de áudio e vídeo, como energia e frequência fundamental do locutor, cor dominante e tipo de tomada de câmera são analisadas. Estas características de áudio e vídeo alimentam um classificador baseado em AdaBoost a fim de identificar os melhores momentos de uma partida de futebol. O sistema gerou resumos que ficaram em torno de 12% do tempo total, contendo 97% dos melhores momentos da partida.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

AUTOMATIC HIGHLIGHTS SUMMARIZATION OF BROADCASTED SOCCER VIDEOS

Luiz Gabriel Lins Bentes Mendonça de Vasconcelos

June/2011

Advisors: Sergio Lima Netto
Eduardo Antônio Barros da Silva

Department: Electrical Engineering

Recent developments in capture, storage and retrieval of video have resulted in a substantial increase in the availability of sports-related multimedia. In addition, content providers have expanded the type and number of sporting events that are available for broadcasting from around the world, in response to a significant increase in audience's interests. Together, these trends point to the necessity for the development of efficient and effective tools to identify highlights in these events and thus reduce viewers' and companies' efforts in searching for material of interest. This dissertation describes an algorithm for automatic detection of event highlights in videos of soccer matches, where highlights are defined as goals and goal attempts that occurred along the match. In the proposed algorithm, audio and video features such as voice energy and fundamental frequency of the speaker, dominant colour and shot-type of the frame are analyzed. These audio and video features feed an AdaBoost-based classifier to identify the highlights of the soccer match. The framework generated match summaries around 12% of total time, containing 97% of the match highlights.

Sumário

Capítulo 1

Introdução

Ao longo dos últimos anos, a constante evolução e diminuição dos custos das tecnologias de captura, armazenamento, processamento e distribuição de mídia têm aumentado significativamente a produção e o consumo de conteúdo contido em vídeos, tais como filmes, esportes, noticiários e entretenimento em geral.

Da mesma forma, com o passar dos anos, transmissões de praticamente todos os eventos esportivos passaram a ser usuais, o que, devido ao grande interesse da audiência por esportes, essencialmente futebol, exigiu maiores investimentos pelas emissoras de TV no tratamento interno dessas transmissões. Além das transmissões propriamente ditas, o seu conteúdo relevante costuma ser bastante utilizado posteriormente em programas secundários, para arquivo e geração de estatísticas, apesar de CHANG [1] afirmar que o valor de eventos esportivos caem drasticamente após a sua exibição ao vivo, pois o resultado da partida já é conhecido.

Este aumento acelerado do conteúdo esportivo produzido pelas emissoras de TV aliado à dificuldade em manuseá-lo nos processos intrínsecos ao funcionamento da emissora trouxeram à tona o problema que motiva este trabalho: Como encontrar o que eu preciso em meio a uma imensa quantidade de dados no menor tempo possível (SMEATON *et al.* [2])? Ao tentar responder esta pergunta para o futebol, que é o caso desta dissertação, a questão é ainda mais crítica uma vez que os momentos de interesse são minoria no tempo total de transmissão.

Por isso, têm sido cada vez mais intensos os esforços empregados pelos pesquisadores da área de aprendizado semântico do conteúdo de vídeos com o intuito de desenvolver ferramentas computacionais que otimizem os processos onde há o tratamento de transmissões esportivas. No contexto deste trabalho, entende-se o termo "semântico" como uma anotação que pode ser feita intuitivamente pelo usuário ao observar o vídeo onde quem desconhece o esporte, como uma máquina, não seria capaz de fazer. Como exemplo deste processo, tem-se a classificação de uma bola entrando na rede como gol, ou um árbitro exibindo um cartão como uma punição, entre diversos outros eventos comuns a uma partida de futebol.

1.1 Superando a Distância Semântica

Quando um usuário utiliza um sistema que se propõe a realizar a análise semântica de vídeos esportivos, ele espera que o sistema tenha a mesma capacidade dos humanos de fazer este tipo análise e determinar o que ocorre no vídeo. Entretanto, como computadores não possuem este poder de compreensão semântica, a solução encontrada foi criar ligações conceituais entre os eventos semânticos desejados e as evidências contidas no vídeo que são comumente identificadas por algoritmos computacionais.

Em geral a identificação dos eventos de um vídeo é alcançada através da extração de características de baixo e alto nível que podem ser de qualquer natureza, tais como áudio, imagem ou texto. São denominadas de baixo níveis as características que não contêm significado semântico e podem ser identificadas diretamente por algoritmos, tais como cor, textura, formato, classificação de tomada de cena, enquanto são denominadas alto nível as características que exigem um mínimo de entendimento das características de baixo nível, tais como as regras cinemáticas empregadas pela produção na transmissão televisiva.

Por exemplo, no caso específico do futebol, o sistema pode detectar um corte de cena seguido de uma cena em câmera lenta, que seriam características de baixo nível, para afirmar que houve uma ação seguida de um *replay*. Assim, o sistema pode inferir que algo de interessante ocorreu antes do corte de cena, que seria uma característica de alto nível. Em outro exemplo, no caso da sinuca, o desaparecimento de uma bola, uma característica de baixo nível, indica que uma bola entrou na caçapa, e consequentemente ocorreu um evento interessante para o sistema.

A análise baseada em conteúdo de vídeos de esportes pode ser abordada das duas maneiras ilustradas na Figura 1.1 e descritas a seguir:

- *Top-Down:* tem como objetivo descrever todos os eventos ocorridos durante a transmissão. Em uma analogia, XIONG *et al.* [3] compara esta abordagem com a tabela de conteúdo que vem no início de um livro, pois todo o seu conteúdo está listado em ordem de ocorrência; e

- *Bottom-Up:* é útil para encontrar em meio à transmissão exatamente o que se quer. Na mesma analogia, esta abordagem é comparada ao índice que vêm ao fim do livro, onde o conteúdo é descrito por evento e não há ordem de ocorrência.

Apesar de utilizarem métodos similares para extração de características, as duas categorias de sistema se diferem essencialmente pelo fato da primeira precisar necessariamente reconhecer e classificar todos eventos ocorridos ao longo do vídeo em análise, enquanto a segunda pode se resumir à tarefa de reconhecer somente o evento

(a) (b)

Figura 1.1: Abordagens *Top-Down* (a) e *Bottom-Up* (b) para organização do conteúdo em sistemas de análise semântica.

que a interessa sem necessariamente identificá-lo e classificá-lo, o que reduziria sua complexidade.

Um caso particular da abordagem *Bottom-Up* é um sistema de sumarização de partidas de futebol, que se restringe a detectar os melhores momentos da partida. Se utilizarmos a mesma analogia feita por XIONG *et al.* [3], um sistema que encontra os melhores momentos da transmissão de futebol pode ser comparado ao resumo do livro, ou até mesmo à capa de uma revista como ilustra a Figura 1.2.

Figura 1.2: Sistema de sumarização de partidas de futebol pode ser comparado à capa de uma revista, onde os principais acontecimentos são destacados.

Outro fator que pode interferir no desenvolvimento do sistema de dada categoria de sistema é o progresso de pontuação do esporte tratado, como ilustrado na Figura 1.3. Em esportes como tênis, beisebol e vôlei, há um vencedor somente quando um dos participantes atinge determinada pontuação, o que torna a utilização de sistemas *Top-Down* adequada, pois há uma sequência definida de fatos que levam aos pontos. Neste caso, é mais difícil definir a importância de um ponto pois eles ocorrem a todo instante.

Por outro lado, em esportes como futebol e futebol americano que são limitados pelo tempo, os eventos de interesse ocorrem aleatoriamente e geralmente com uma frequência muito menor que nos esportes orientados à pontuação. Isto faz com que haja uma dificuldade maior em estruturar as ocorrências analisadas de forma a se encaixar na categoria *Top-Down*.

Assim, a abordagem orientada pelo tempo, que é o caso do futebol, torna bastante interessante o uso de sistemas *Bottom-Up*, pois em geral na maior parte do tempo não há nada de interessante acontecendo. Dessa forma, este trabalho idealiza a construção de um sistema de sumarização para o futebol, onde somente as pontuações e as ameaças de pontuação serão destacadas. Para isso, o entendimento

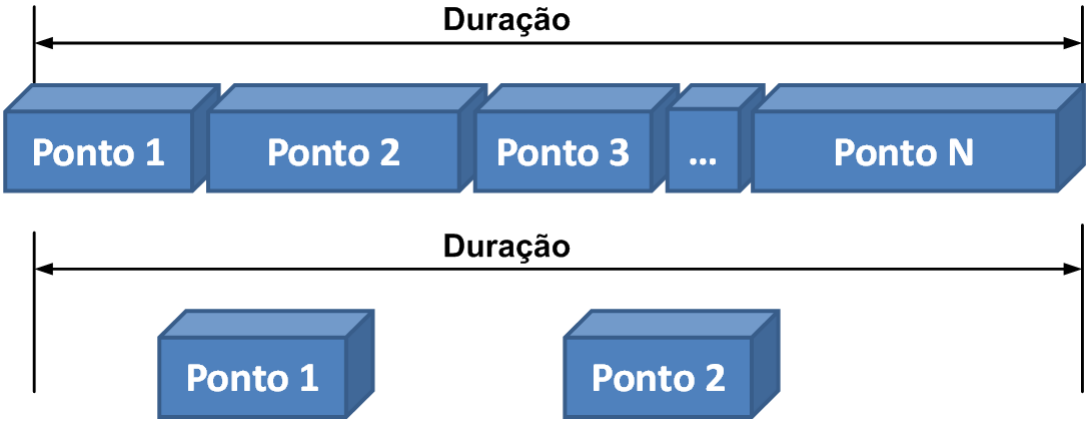


Figura 1.3: Em esportes orientados à pontuação, a duração não é determinada e pontuações ocorrem uma seguida da outra. Já em esportes orientados por tempo, a duração é previamente determinada e as pontuações ocorrem aleatoriamente, podendo até não acontecer.

semântico da transmissão de uma partida de futebol será baseado em um classificador que combinará diversas características audiovisuais.

Vale ressaltar que o sistema proposto neste trabalho destina-se exclusivamente a vídeos obtidos de transmissões de televisão, pois as regras de produção empregadas pela emissora de TV, tais como locução, mudança de câmeras entre outros, são essenciais para o correto funcionamento do sistema.

1.2 Organização da Dissertação

O Capítulo 2 explorará a atuação da voz do locutor contida no áudio da transmissão da partida de futebol para extrair características que possam contribuir para o sistema de sumarização proposto neste trabalho. Neste capítulo, ainda haverá uma breve revisão do que tem sido feito a respeito na literatura, e, posteriormente, serão feitas análises sobre os algoritmos propostos por este trabalho para extração de características baseadas na frequência fundamental e energia da voz do locutor.

O Capítulo 3 também fará uma breve revisão de como o vídeo tem contribuído para sistemas de sumarização ao longo da literatura. Após isto, serão estudados e propostos algoritmos de análise da cor dominante de cada quadro de vídeo e do movimento de câmera ocorrido durante a transmissão que resultarão em características associadas ao vídeo.

Em seguida, o Capítulo 4 mostrará, em primeiro lugar, como trabalhos deste tipo realizam a combinação de características multi-modais na literatura. Com isso, será proposta uma metodologia de classificação baseada no classificador *AdaBoost* (*Adaptive Boosting*) que combinará as características extraídas do áudio com as extraídas do vídeo. Também serão propostos pré- e pós-processamentos ao *AdaBoost*

para adaptá-lo às necessidades de nosso sistema de sumarização. Ainda neste capítulo, tarefas de validação e experimentos serão discutidos a fim de definir um método robusto de avaliação desempenho do sistema.

Após isto, o Capítulo 5 definirá o sistema, para, depois, realizar o treinamento, as validações e os experimentos propostos no Capítulo 4. Assim, será possível colocar em números o desempenho do sistema desenvolvido ao longo da dissertação, para posterior análise.

Finalmente, o Capítulo 6 faz uma retrospectiva de toda a dissertação, com comentários a respeito dos resultados obtidos, contribuições do projeto e propostas para trabalhos futuros.

Capítulo 2

Características de Áudio

O objetivo principal deste capítulo é descrever as características de áudio que possam ser interessantes para um sistema de sumarização de transmissões esportivas. A Seção 2.1 fará uma breve revisão do que tem sido feito na literatura associada para utilizar o áudio neste tipo de sistema. Depois, a Seção 2.2 apresentará os algoritmos propostos por este trabalho para a extração de características baseadas na frequência fundamental e energia da voz. Por fim, a Seção 2.3 comentará as análises realizadas ao longo deste capítulo e fará as considerações finais.

2.1 Revisão Bibliográfica

Na literatura relacionada à sumarização e recuperação de informações de vídeos de esportes, encontram-se diversas referências que exploram variadas características do áudio de diferentes maneiras. Esta seção descreverá alguns destes trabalhos, organizados por tipos de características.

Um dos pontos onde o processamento de áudio em transmissões esportivas pode ser útil é na identificação de sons específicos, como o som do chute na bola, bola batendo na trave e apito, como proposto em TJONDRONEGORO *et al.* [4]. É comum ver este tipo de abordagem para o beisebol, como em RUI *et al.* [5], e tênis, como em DAHYOT *et al.* [6], em que se detecta o impacto da bola a partir do sinal de áudio. Entretanto, este processamento costuma ser extremamente complexo, exigindo alto custo computacional. Além disso, é necessário confiar que em todas as transmissões esportivas haverá captação de áudio suficiente para tais eventos, o que não é comum em partidas de futebol.

Outras propostas, de XIONG *et al.* [3] e NEPAL *et al.* [7], baseiam-se na identificação de reações da audiência presente no estádio para sugerir que um lance de interesse esteja acontecendo. Esta técnica, entretanto, está diretamente atrelada a diversos fatores que variam bastante de uma partida para outra, tais como o comportamento de determinada audiência fazendo mais ou menos barulho que outra,

se a equipe que marcou um gol está atuando no estádio do time adversário, se a audiência teve uma reação devido ao resultado de um rival que está jogando em outro lugar, etc.

CHANG *et al.* [8] e VOJKAN *et al.* [9] realizam detecções de palavras-chave para determinado evento esportivo, como, por exemplo, a palavra "gol" para o futebol. Apesar de interessante, este tipo de técnica exige treinamento e classificação específicas para reconhecimento de palavras, aumentando significativamente a complexidade para obtenção de uma característica que não necessariamente será determinante na identificação de eventos.

Além destas, as técnicas baseadas em áudio mais exploradas na identificação de lances de interesse são a energia e a frequência fundamental da voz do narrador. Dentre os trabalhos nessa área, podemos citar CABASSON e DIVAKARAN [10], HANJALIC [11], DAGTAS e ABDEL-MOTTALEB [12], REA *et al.* [13] e XIONG *et al.* [3] que utilizam energia em seus sistemas. LEONARDI *et al.* [14] faz uso das *Hidden Markov Models* da energia do áudio para, juntamente com o vídeo, classificar lances como gol e escanteio, dentre outras quatro classes.

Alguns autores, como ZHANG e ELLIS [15] e RUI *et al.* [5] consideram que para que os cálculos de energia e *pitch* sejam bem-sucedidos é essencial identificar a fonte do sinal, e para isto, fazem uso dos coeficientes *Mel-frequency cepstral* e seus derivativos.

EKIN [16] também considerou a identificação da fonte de sinal importante, mas somente para o cálculo do *pitch*. Com o intuito de criar um sistema em tempo-real, o autor desprezou o *pitch* e propôs um sistema causal, ou seja, onde somente são utilizadas amostras disponíveis no instante da captura, para encontrar picos de energia. Dessa forma, diferentemente de outros trabalhos que utilizam estatísticas globais de energia para definição dos limiares, este trabalho define os mesmos de acordo com as estatísticas das amostras disponíveis naquele momento.

Em uma tentativa de criar um sistema não-supervisionado, ou seja, que não contenha estágio de treinamento, COLDEFY e BOUTHEMY [17] definiram algoritmos que após extrair as características de energia e *pitch*, fossem capazes de inferir os melhores momentos ao detectar o crescimento dessas características ao longo do tempo.

Por fim, VASCONCELOS *et al.* [18], em um trabalho preliminar a este, construíram uma abordagem, que com base em análises simples de energia e *pitch*, se propôs a encontrar os melhores momentos de uma partida de futebol. Para o mesmo narrador utilizado no treinamento, este sistema foi capaz de conter 100% dos melhores momentos de cada partida em um conjunto de lances com uma taxa de 50% de falsos positivos. Entretanto, para narradores diferentes ao utilizado no treinamento, a taxa de acerto foi bem abaixo do esperado.

Apesar de utilizar o áudio em seus sistemas, alguns autores afirmam que, usualmente, o áudio é usado de forma suplementar na definição da ocorrência de um momento de interesse em transmissões esportivas. Porém, assim como CHANG *et al.* [8], RUI *et al.* [5], VOJKAN *et al.* [9] e COLDEFY e BOUTHEMY [17], nós acreditamos que as características de áudio podem ser tratadas como determinantes para a identificação de melhores momentos. E, para isso, pode-se extrair características baseadas na voz do locutor da partida, uma vez que ele tem a responsabilidade de excitar a sua voz nestes momentos em transmissões esportivas, como afirma OWENS [19].

2.2 Características de Áudio

Segundo COWIE *et al.* [20], GERHARD [21] e ROCCHESSO [22], pode-se caracterizar momentos de excitação da voz por valores altos de energia e *pitch*. Isto torna interessante a análise das variações destas características ao longo do tempo. Por este motivo, este trabalho terá as características de áudio baseadas nas análises de *pitch* e energia da voz, sendo que estas foram inspiradas nos algoritmos propostos por VASCONCELOS *et al.* [18] e COLDEFY e BOUTHEMY [17].

2.2.1 Estimação de *Pitch*

Para a estimação do *pitch*, a idéia principal é explorar a periodicidade presente em trechos sonoros dos sinais de voz. Dentre os métodos de estimação de *pitch* propostos em GERHARD [21], o operador de auto-correlação R_{xx} explora esta característica. Assim, para um sinal de áudio $x(n)$ tem-se

$$R_{xx}(\tau) = \sum_{n=1}^{N} x(n)x(n-\tau) \quad . \tag{2.1}$$

Um exemplo desta função é indicado na Figura 2.1 de onde é possível extrair o período T, e consequentemente o seu inverso, que é a frequência fundamental F_0 do sinal de áudio.

O uso da auto-correlação para estimação do *pitch* torna-se ainda mais interessante se considerarmos que TOLONEN e KARJALAINEN [23] propõem o cálculo no domínio da frequência com o objetivo de atingir melhor desempenho computacional. Assim, a auto-correlação passa a ser obtida a partir de

$$R_{xx}(\tau) = IDFT\{|DFT[x(n)]|^2\} \quad , \tag{2.2}$$

onde $DFT(.)$ e $IDFT(.)$ são as transformadas discretas de Fourier direta e inversa, respectivamente.

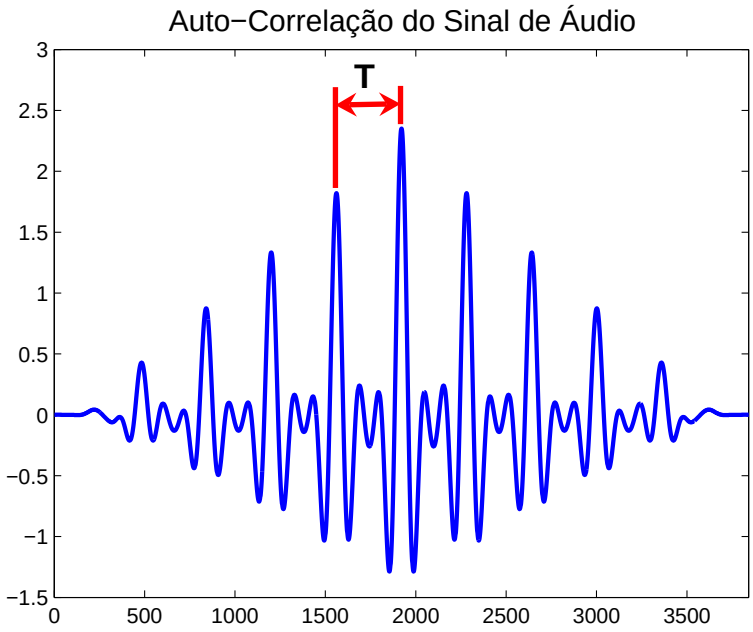


Figura 2.1: Estimação do período de *pitch* do sinal de áudio.

Em geral, o sinal de áudio de uma transmissão esportiva de TV, essencialmente o futebol, não contém somente a voz do locutor, pois também possui o áudio ambiente. Apesar de alguns trabalhos apresentarem técnicas para identificação da fonte sonora, é possível reduzir as influências de fontes não provenientes da voz somente aplicando algumas simples medidas:

- A primeira medida é assumir que valores de *pitch* não usuais sejam descartados. Portanto, seguindo ROCCHESSO [22], este trabalho assume que qualquer valor de *pitch* abaixo de 50 Hz e acima de 500 Hz ocorrem raramente em sinais de voz.

- Outra medida interessante é também anular o cálculo do *pitch* em trechos onde não há energia suficiente para um trecho sonoro, isto é, caracterizando um silêncio.

A Figura 2.2 mostra um sinal de áudio aplicado aos métodos de estimação de *pitch* proposto nesta dissertação, e o por COLDEFY e BOUTHEMY [17]. A diferença entre o proposto por este trabalho e o proposto por COLDEFY e BOUTHEMY [17] está no fato de que o primeiro evita o cálculo do *pitch* quando da detecção de silêncio. É fácil notar que em ambos os métodos, o gol contido no trecho entre 20 e 25 segundos visivelmente se destaca dos demais, já que o valor do *pitch* permanece mais alto.

Por se tratar de um sistema multi-modal, é vantajoso ajustar a taxa de amostragem do áudio para que fique compatível com a taxa de quadros do vídeo, ou seja,

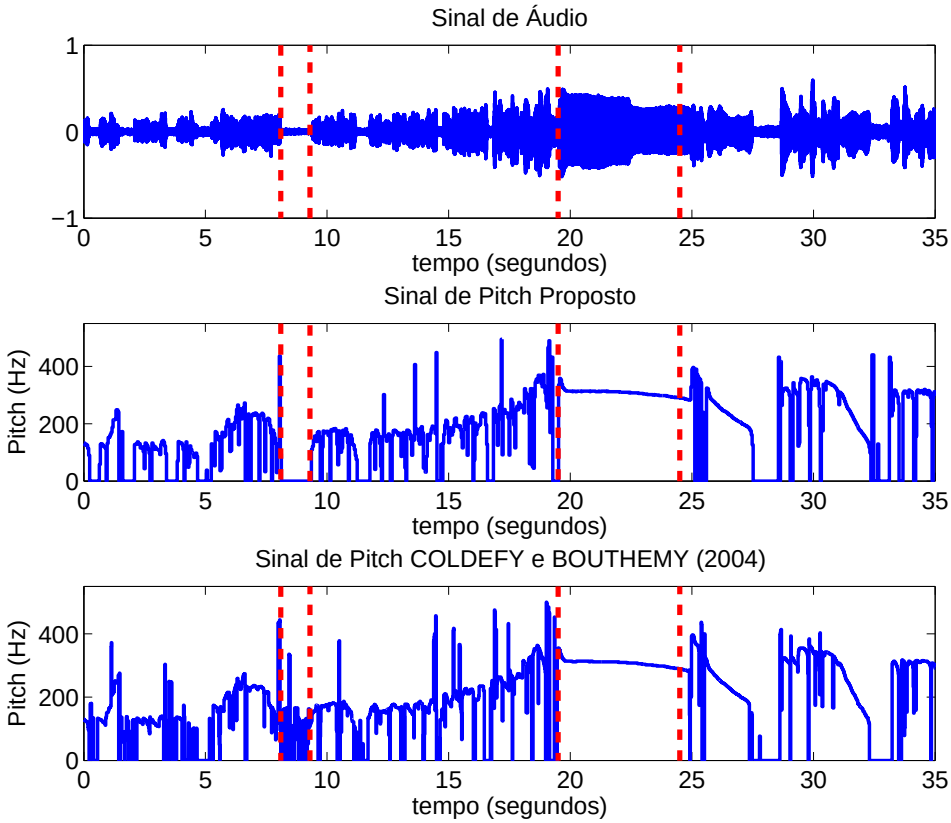


Figura 2.2: Sinal de áudio e estimação de *pitch* proposta por este trabalho ao longo do tempo junto com a estimação de *pitch* de COLDEFY e BOUTHEMY [17]. Com o método proposto, o silêncio é melhor caracterizado.

passar dos usuais 48000 Hz do áudio para os 29.97 quadros por segundo do sistema de vídeo NTSC, que será usado neste trabalho. Assim, com o intuito de obter um único valor de *pitch* para cada segmento de voz, este trabalho usará janelas consecutivas de 33.3 milissegundos que é aproximadamente o tempo de um quadro de vídeo.

2.2.2 Medida de Energia

Quando fala-se em análise da energia de voz, usualmente fala-se em calcular a energia total de uma janela aplicada ao sinal de áudio. Este método é denominado Energia Tempo-Curto e definido por

$$e_{st}(t) = \sum_{n=t-\frac{N_h}{2}}^{t+\frac{N_h}{2}} x^2(n)h(n) \quad , \tag{2.3}$$

onde $x(n)$ é o sinal de áudio filtrado, que será explicado posteriormente, e N_h é o tamanho da janela $h(n)$ que será aplicada ao sinal de áudio no domínio do tempo t.

Esta janela pode ser retangular, *Hamming*, *Hanning* entre outros tipos encontrados em DINIZ *et al.* [24].

Mais uma vez, para analisar a voz a partir de um sinal de áudio contendo outras fontes sonoras, é interessante enfatizar a voz do locutor. Com este objetivo, um filtro passa-baixas foi utilizado antes do cálculo da energia reduzindo as componentes de alta frequência comumente associadas a ruído.

Além disso, aplicamos um filtro em pente (*Comb Filter*) no domínio da frequência para enfatizar mais ainda a energia que pertence à voz do locutor. Neste filtro, a idéia é calcular somente a energia contida em bandas que estão ao redor da frequência fundamental e seus harmônicos. Para o projeto do filtro *Comb*, utiliza-se o método de estimação de *pitch* empregado na Seção 2.2.1. Isto é feito em uma janela de tamanho N_h no domínio do tempo, que terá o espectro calculado, para, posteriormente, a energia e_{cs} ser calculada utilizando o filtro *Comb* C no domínio da frequência, definido por

$$C(f) = \sum_{k=1:d} \delta(f - kF_0) * \Pi(f) \quad , \tag{2.4}$$

onde δ é a função impulso de *Dirac*, F_0 é o *pitch* estimado, d o número de harmônicos que terão suas energias computadas e Π é a janela do filtro *Comb*, que pode ser de qualquer tipo, da mesma forma que a janela do cálculo de energia tempo-curto apresentada anteriormente.

A Figura 2.3 mostra um exemplo do filtro *Comb* aplicado ao espectro do sinal de áudio, onde é possível observar que em frequências que não estão próximas ao *pitch* estimado e seus harmônicos, a energia e_{cs} não será computada. Além disso, quando o *pitch* estimado estiver fora da faixa considerada de voz, estipulada na Seção 2.2.1, a energia e_{cs} também não será computada.

Por fim, pode-se ainda propor uma pequena modificação para um novo cálculo de energia, que foi denominado como Soma *Comb* de Energia Tempo-Curto Modificada e_{mcs}. Esta modificação consiste em aplicar a primeira derivada no sinal de áudio antes do filtro passa-baixas, o que é avaliado como uma evolução da medida de energia para vozes excitadas. A adição deste operador é motivada pelo fato de que em momentos de voz excitada, a energia costuma ter um aumento significativo nas frequências mais altas. Em resumo, esta medida simplesmente visa valorizar a energia em altas frequências de voz.

Os valores dos parâmetros utilizados nestas abordagens de energia serão os mesmos que os propostos por COLDEFY e BOUTHEMY [17]. Assim, será utilizada a janela de *Hamming* para a aplicação da DFT no domínio do tempo, a janela $\Pi(f)$ será do tipo retangular de largura 50 Hz no domínio da frequência na utilização do filtro *Comb*, com $k = 10$ harmônicos na Equação (2.4) e o filtro aplicado antes do

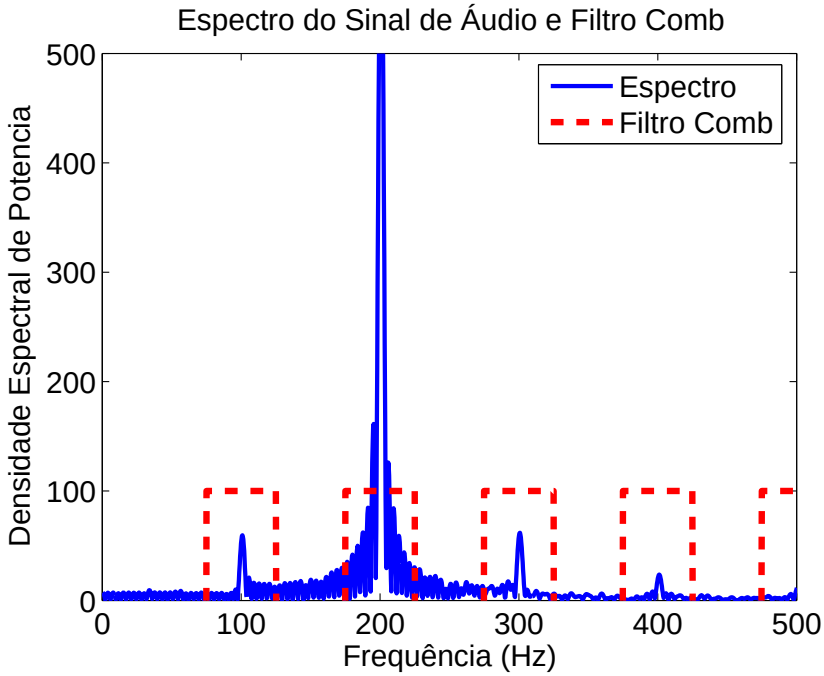


Figura 2.3: Sinal de áudio no domínio da frequência e filtro *Comb* aplicado para cálculo da energia somente nos arredores da frequência fundamental e seus harmônicos.

cálculo da energia terá 4400 Hz como frequência de corte. Somente o tamanho da janela de *Hamming* será de 33 milissegundos para sincronizar o áudio com vídeo, assim como para o caso da estimação de *pitch* explicado na Seção 2.2.1.

A Figura 2.4 mostra exemplos de um sinal de áudio aplicado às três abordagens descritas nesta seção. É possível notar que de maneira geral as três abordagens são capazes de diferenciar o trecho de gol, que ocorre entre 20 e 25 segundos aproximadamente, dos trechos anterior de locução normal e posterior de repercussão do gol. Entretanto, a abordagem de Energia Tempo-Curto apresenta variações mais intensas, enquanto as baseadas no filtro *Comb* apresentam sinais de energia mais suaves e que melhor destacam o trecho do gol.

Apesar de visualmente as técnicas baseadas no filtro *Comb* aparentarem ser mais eficientes na identificação de momentos de interesse, todas as características serão exploradas pelo classificador que será apresentado no Capítulo 4, deixando para o próprio classificador escolher quais são úteis.

2.2.3 Detecção de Crescimento Local

Uma vez definidos os algoritmos de estimação do *pitch* e de medida da energia nas Seções 2.2.1 e 2.2.2, respectivamente, o próximo passo é encontrar uma maneira de inferir quando algo interessante acontece na partida através destes sinais. Na prática, a idéia desta seção é entendida como uma forma de detectar quando há aumento

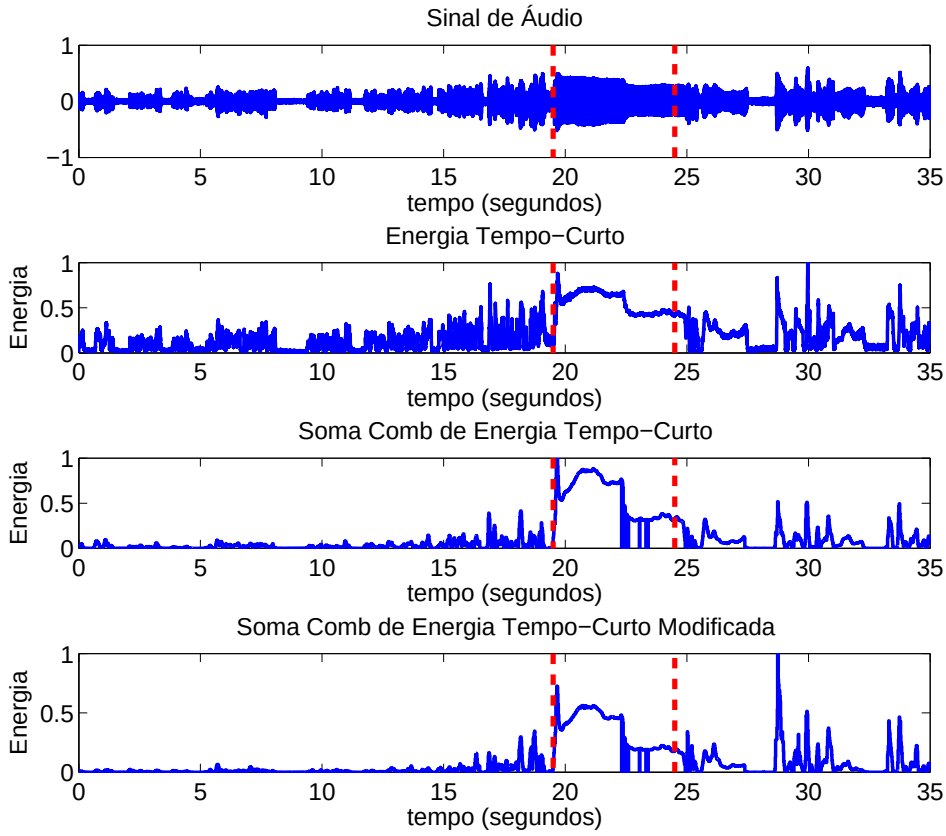


Figura 2.4: Sinal de áudio, Energia Tempo-Curto, Soma *Comb* de Energia Tempo-Curto e Soma *Comb* de Energia Tempo-Curto Modificada. Os métodos baseados no filtro *Comb* destacam melhor o evento do gol.

significativo no valor destas características, já que isto indicaria que o locutor está aplicando mais excitação à sua voz do que em amostras anteriores.

A princípio, sempre se pensa em determinar limiares que, quando ultrapassados, indicarão que aquele momento pode ser considerado interessante. Entretanto, limiares fixos não são ideais para sistemas que contêm variações de locutores (frequências fundamentais e volumes da voz diferentes, principalmente se houver um locutor do sexo feminino), ou do equipamento de captação de áudio (que por algum motivo, que está fora do escopo deste trabalho, pode variar o volume do áudio de um mesmo locutor de uma transmissão para a outra, ou até na mesma transmissão).

Para evitar efeitos destas possíveis variações, podemos calcular as médias μ_a e μ_p das janelas de N_a amostras anteriores e N_p amostras subsequentes à amostra t que está sendo avaliada. Este cálculo pode ser aplicado ao longo do sinal contendo a evolução no tempo de cada uma das características $f(n)$ por meio de

$$\mu_a(t) = \sum_{n=t+1}^{t+N_a} f(n) \quad , \tag{2.5}$$

$$\mu_p(t) = \sum_{n=t-N_p}^{t-1} f(n) \quad . \tag{2.6}$$

Uma vez calculadas as médias μ_a e μ_p, basta, então, computar a magnitude μ_m da diferença $\mu_m(t) = |\mu_p(t) - \mu_a(t)|$ que indica se houve crescimento ou decrescimento desta característica, através da relação

$$\mu_s(t) = \begin{cases} 1, & \text{se } \mu_p(t) - \mu_a(t) \geq 0 \\ 0, & \text{se } \mu_p(t) - \mu_a(t) < 0 \end{cases} \quad . \tag{2.7}$$

A Figura 2.5 compara a marcação baseada em limiar com o método de detecção de crescimento local por diferença demonstrado neste trabalho para o caso de informação de *pitch*. Neste exemplo, foi utilizado o limiar de 225 Hz para o primeiro caso e 50 Hz para o segundo.

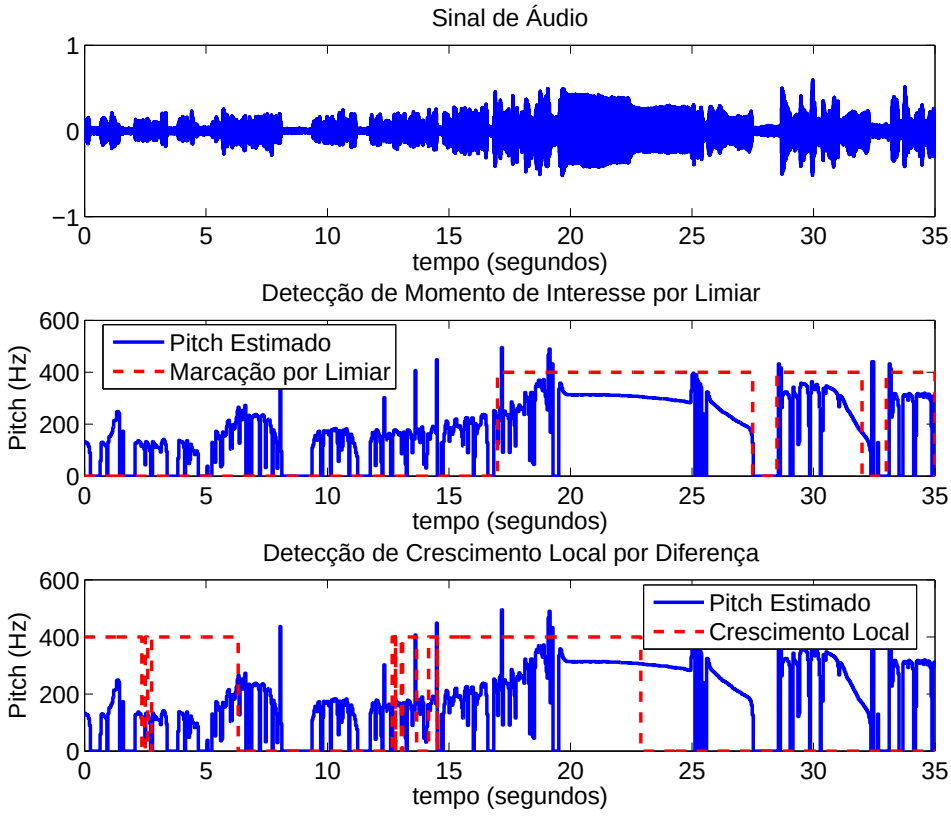


Figura 2.5: Sinal de áudio e marcação de momentos de interesse para a detecção por limiar absoluto e por diferença entre janelas, respectivamente. O método baseado em limiar detecta altos valores de *pitch*, enquanto o outro detecta crescimentos.

É possível notar que ambos os métodos identificam o gol na Figura 2.5. Entretanto, também nota-se que enquanto o baseado em limiar marca o momento de gol inteiro e sua repercussão, o baseado em diferença detecta somente o início do

gol onde há o visível crescimento do valor de *pitch*, como era de se esperar. Apesar de parecer marcar melhor um momento de interesse, a abordagem que utiliza limiar depende exclusivamente de um valor absoluto que pode variar consideravelmente de um locutor para outro e de uma transmissão para outra, como já foi dito anteriormente. Portanto, a detecção de crescimento por diferença de médias entre janelas apresenta-se como uma solução para generalizar a detecção de momentos de interesse para este trabalho.

Com o objetivo de generalizar também a medida de energia da Seção 2.2.2, a detecção de crescimento local já demonstrada também foi aplicada para este caso. Mais uma vez, somente para exemplificar o efeito do método de detecção de crescimento local na medida de energia, foi definido o limiar 0.1. A Figura 2.6 exibe este exemplo.

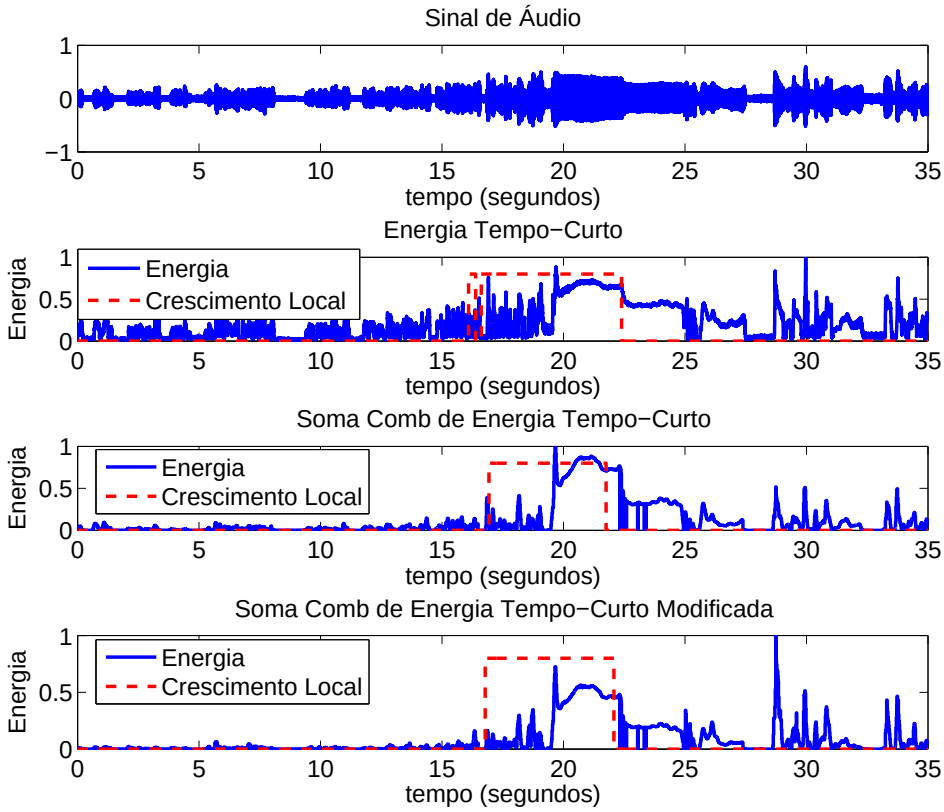


Figura 2.6: Sinal de áudio, Energia Tempo-Curto, Soma *Comb* de Energia Tempo-Curto, Soma *Comb* de Energia Tempo-Curto Modificada e suas respectivas marcações de momentos de interesse de acordo com a detecção de crescimento local. O algoritmo foi capaz de detectar o aumento de energia no momento do gol para as três abordagens.

Novamente o algoritmo foi capaz de identificar o momento em que o gol ocorreu em todas as três abordagens de energia. Com isso, fica claro que o algoritmo de detecção de crescimento local é bem-sucedido tanto para o *pitch* quanto para a energia.

Os tamanhos das janelas anterior e posterior valendo 10 e 3 segundos, respectivamente, foram definidos de acordo com o estabelecido por COLDEFY e BOUTHEMY [17].

2.3 Conclusões

O objetivo deste capítulo foi estudar as características do áudio de uma transmissão televisiva de futebol com a finalidade de extrair características que possam ser úteis ao classificador de momentos de interesse desta transmissão.

A Seção 2.1 realizou uma breve revisão bibliográfica onde foi possível verificar o que tem sido feito com as características de áudio para que estas sejam significativas em sistemas de identificação de momentos de interesse.

Em sequência, a Seção 2.2.1 propôs um método de estimação de *pitch* baseado no cálculo de auto-correlação já utilizado no trabalho preliminar a este, porém com alguns tratamentos que têm por objetivo reduzir a interferência de fontes que não são de voz. Dentre eles, está a suposição de que qualquer trecho com valor de *pitch* estimado fora da faixa de 50 a 500 Hz não é voz.

Após isso, a Seção 2.2.2 discutiu três abordagens baseadas na Energia Tempo-Curto de uma janela ao longo do sinal de áudio e a aplicação de um filtro *Comb* no domínio da frequência, que também têm como objetivo reduzir a interferência de fontes que não são provenientes da voz.

Por fim, a Seção 2.2.3 descreveu um algoritmo capaz de identificar crescimentos locais que tem a finalidade de generalizar o sistema, reduzindo o impacto das variações inerentes às mudanças de locutor, equipamentos e configurações na captura de áudio que ocorrem usualmente nas transmissões de futebol. A ideia basicamente consiste em calcular a diferença de médias entre janelas anterior e posterior a cada amostra.

Em resumo, este capítulo gerou doze características para o classificador, sendo três relativas à frequência fundamental e nove relativas à energia. Estas características de áudio posteriormente serão combinadas com as características de vídeo, que serão discutidas no próximo capítulo, para serem utilizadas no classificador que será apresentado no Capítulo 4.

Capítulo 3

Características de Vídeo

Uma vez descritas as características de áudio no Capítulo 2, o presente capítulo trata das características de vídeo que possam ser interessantes a sistemas de sumarização de vídeos oriundos de transmissões esportivas. Em primeiro lugar, a Seção 3.1 mostrará a importância das características de vídeo assim como os algoritmos já desenvolvidos para sua extração neste tipo de sistema. Em seguida, na Seção 3.2 serão apresentadas as características e seus respectivos algoritmos utilizados neste trabalho. Por fim, a Seção 3.4 resumirá este capítulo e realizará as considerações finais pertinentes.

3.1 Revisão Bibliográfica

Tipicamente, há dois grupos de sistemas de classificação de conteúdo de vídeo: O primeiro é composto por sistemas multi-modais que analisam vídeo, áudio e texto obtidos através de uma transmissão televisiva em busca de características de cor, movimento de câmera, objetos presentes, entre outros, para classificar determinado evento. O segundo grupo é composto por sistemas que utilizam múltiplas câmeras controladas pelo sistema, para definir as posições de todos os objetos importantes para determinado esporte para, a partir disto, indicar o que está ocorrendo a cada instante. Assim, nesta seção serão revistas as características de vídeo mais relevantes para este trabalho, apenas citando algumas outras características que também aparecem na literatura.

3.1.1 Cor Dominante

A aplicação mais comum no uso de cor dominante encontrada na literatura tem o intuito de identificar e classificar o campo de jogo dos diversos esportes. SUDHIR *et al.* [25] calculam a cor dominante usando o espaço de cor RGB (vermelho-verde-azul) da área central dos primeiros quadros de vídeos para inferir o tipo da quadra de

tênis em questão, que podem ser saibro, piso rápido ou grama. De forma similar, LI e SEZAN [26] também calculam a cor dominante da área central dos quadros iniciais do vídeo em partidas de futebol americano para identificar se dado quadro é do jogo. Esse trabalho, porém, utilizou o espaço de cores HSV (matiz-saturação-valor) e procurava identificar a formação característica de jogadores do futebol americano no início de cada jogada para atualizar a cor dominante utilizada como referência.

ELDIB *et al.* [27] utilizam o espaço de cor RGB e limiares pré-definidos para cada componente com o intuito de segmentar a região do quadro de vídeo que contém o campo de jogo para o caso do futebol. Porém, estes limiares nem sempre são adequados devido às diversas variações entre os estádios que influenciam na cor do campo.

Outra técnica bastante utilizada é considerar como cor dominante os picos dos histogramas das componentes do espaço de cor utilizado em cada trabalho. SEO *et al.* [28] calculam a distância das componentes RGB de cada *pixel* para os picos dos histogramas, que são assumidos como verde, para determinar quais *pixels* são pertencentes ao campo efetivamente. Esta técnica falha quando o quadro não contém o campo de futebol, pois os picos dos histogramas não serão relativos à cor verde. Da mesma forma, XU *et al.* [29] e REA *et al.* [13] também obtêm a cor dominante através de histogramas, porém utilizando outros espaços de cores.

Em outro trabalho, EKIN [16] propôs um algoritmo para todos os esportes robusto a variações de cor, tais como diferenças de cor no mesmo campo, mudanças climáticas, variações na iluminação e variações no campo ao longo do tempo. Esta robustez é obtida através da utilização de dois espaços de cores complementares na análise de histogramas do quadro de vídeo.

COLDEFY e BOUTHEMY [17] também abordam histogramas de componentes de cores, mas para aplicar a um algoritmo k-médias que dirá se o centróide encontrado para cada quadro está próximo à cor verde ou não.

DAGTAS e ABDEL-MOTTALEB [12] detectam a presença de regiões de campo de futebol para simplesmente afirmar que o quadro em questão é de um campo de grama ou não. Desta vez, são utilizados histogramas das componentes do espaço de cor YCbCr (luminância e crominâncias).

Para casos onde há mais de uma cor dominante no campo de jogo, como o beisebol, HUNG e HSIEH [30] propuseram um algoritmo de segmentação do campo através de estatísticas dos histogramas das componentes de cor que são automaticamente ajustadas ao longo do tempo para suportar as mudanças constantes de cor dominante.

Por fim, NGO *et al.* [31] fazem uso de redes GMM (*Gaussian Mixture Models*) para estimar grosseiramente a região do campo de jogo, para posteriormente aplicar o conceito de área-de-cobertura de uma câmera para refinar o resultado. Este conceito

nada mais é que do que tentar inferir o ângulo da câmera em relação ao campo, para a partir disto adicionar à região do campo de jogo os *pixels* que as GMMs não foram capazes de adicionar.

Pelo fato da cor dominante ser uma característica de baixo-nível, ela é bastante utilizada para extração de outras características que serão vistas ao longo desta seção, que envolvem segmentação e classificação de cenas.

3.1.2 Movimento de Câmera

O movimento das câmeras de uma determinada transmissão esportiva está diretamente atrelado à forma como as equipes de televisão produzem este evento. EKIN [16] afirma que existem técnicas de produção que são regularmente utilizadas por estas equipes, o que torna as características extraídas do movimento das câmeras e sua evolução temporal bastante úteis para sistemas de sumarização e recuperação de informações em transmissões esportivas.

Por exemplo, alguns trabalhos, como SAUR *et al.* [32], SAUR *et al.* [33], ZHOU *et al.* [34] e NEPAL *et al.* [7], utilizam a direção e sentido do movimento dominante para inferir que um ataque está sendo iniciado para o caso do basquete, além de alguns outros eventos, tais como roubadas de bola e arremessos perdidos.

Entretanto, os parâmetros de movimento dominante de câmera são mais utilizados como base para algoritmos que exploram outras características de vídeo, tais como detecção de início, fim e classificação de cenas de vídeo, que serão vistas com maior profundidade nas próximas seções. Neste contexto, LAZARESCU *et al.* [35] e CHANG *et al.* [36] utilizaram parâmetros de movimento de câmera para classificação de cenas para o *cricket* e beisebol.

Uma evolução deste tipo de caracterização é a análise dos parâmetros ao longo do tempo. COLDEFY e BOUTHEMY [17] assumem que o movimento da câmera panorâmica em partidas de futebol pode ser representado por um modelo afim bidimensional e, assim, identificar os limites das cenas ao longo do tempo.

Alguns trabalhos propuseram HMMs (*Hidden Markov Models*) usando parâmetros de movimento de câmera juntamente com parâmetros de outras naturezas com o intuito de modelar as técnicas de produção empregadas pelas equipes de TV. Por exemplo, LI e SEZAN [37] e XIE *et al.* [38] utilizam HMMs a fim de discriminar tanto sequências de jogo em andamento quanto paradas para o beisebol, futebol e sumô. Já ASSFALG *et al.* [39] combinam parâmetros de movimentação de câmera com a posição dos jogadores no campo para treinar os HMMs. Finalmente, LEONARDI *et al.* [14] utilizam movimentação de câmera com transições de cenas para encontrar eventos tais como o gol em uma transmissão de futebol.

Outros trabalhos, como KOKARAM e DELACOURT [40] para o *cricket*, explo-

ram o deslocamento dos objetos nos quadros de vídeos subsequentes para extrair parâmetros do movimento dominante da câmera.

Há ainda outros trabalhos, tais como PEKER *et al.* [41] e CABASSON e DIVAKARAN [10], que analisam vetores de movimento MPEG diretamente do *bitstream* dos vídeos codificados como uma forma alternativa de análise do movimento de câmera, além de obter melhor desempenho uma vez que a decodificação de vídeo torna-se desnecessária.

3.1.3 Detecção de Cortes de Cena

Apesar de técnicas de detecção de cortes de cenas genéricas poderem ser utilizadas em transmissões esportivas, há diversos trabalhos que analisam as características dos esportes para tratarem o problema especificamente em transmissões esportivas. Para vídeos esportivos, cortes de cena sempre significam os momentos em que ocorre a troca de uma câmera para outra.

Como visto na Seção 3.1.2, muitos trabalhos usam parâmetros provenientes do movimento de câmera para detecção de cortes de cenas. Porém, também são comuns métodos que utilizam informações de cor para tal tarefa. De fato, SAUR *et al.* [33], BABAGUCHI *et al.* [42], EKIN [16] e HUNG e HSIEH [30] propuseram algoritmos diferentes que exploram histogramas de cores em quadros de vídeo adjacentes para realizar a detecção dos cortes de cena. FACON e TEIGÃO [43] propuseram o conceito de ritmo visual, que consiste em aplicar uma transformação do vídeo para uma imagem 2D, onde é possível notar os limites das cenas.

No entanto, algoritmos que consideram somente a análise de cor não são capazes de identificar transições de cena graduais. Assim, para superar este problema, LI *et al.* [44] agregaram uma análise temporal entre duas cenas para detectar transições desta natureza. Em outra abordagem do problema, REFAEY *et al.* [45] utilizaram lógica *Fuzzy*, que, após treinamento, pôde detectar tanto transições em corte seco quanto graduais.

Em outra forma de abordagem, BABAGUCHI *et al.* [46] e PAN *et al.* [47] se baseiam nas inserções e remoções de logos e artes de placar postos em cenas de transmissões televisivas com o objetivo de detectar transições de cenas. Estas técnicas são efetivas para identificar quando *replays*, por exemplo, ocorrem, pois é comum haver transições graduais com inserções de logos nestes casos. Apesar de serem efetivos, estes métodos requerem conhecimento prévio dos logos e artes utilizadas nas transmissões de cada emissora.

3.1.4 Classificação de Cenas

Uma das características mais exploradas na literatura é a classificação de cenas, que basicamente consiste em apontar o que determinado *shot* está exibindo. Mais especificamente para o caso de transmissões esportivas, podemos classificar as imagens como: panorâmicas, onde o campo de ação do esporte é visto de maneira global; médias, onde uma região do campo de ação é destacada; fechadas, onde geralmente pessoas e objetos são os destaques; e outras imagens, onde, por exemplo, a audiência é exibida.

Mais uma vez, são muitos os trabalhos que se baseiam em análise de cores para extrair informações, principalmente os que classificam cada quadro de vídeo isoladamente sem utilizar informações temporais. KAWASHIMA *et al.* [48] extraíam características relacionadas à cor para gerar *templates* de cada tipo de imagem e posteriormente realizar a classificação de novos quadros de vídeo. Em outra forma de abordagem, SUDHIR *et al.* [25] e XU *et al.* [29] calculam quantos dos *pixels* do quadro de vídeo pertencem ao campo de ação para determinar se as imagens são panorâmicas. Já LI e SEZAN [26] combinam cor, textura e formato para também encontrar imagens panorâmicas.

Ainda no âmbito de classificações do quadro sem dependência temporal, CHANG *et al.* [8] e DAHYOT *et al.* [6] realizam classificações de imagens panorâmicas baseadas em algoritmos de detecção de linhas do campo de ação para o futebol e tênis, respectivamente. HUNG e HSIEH [30] utilizam clusterização (k-médias) para detectar cenas de *pitch* (região de onde a bola é arremessada) e campo para transmissões de beisebol.

Apesar dos trabalhos já citados se apresentarem eficazes na classificação de cenas, muitos autores acreditam que é interessante investigar características que variam ao longo do tempo, como o movimento de câmera (já visto na Seção 3.1.2), e movimento de objetos. Com isso, ZHONG e CHANG [49] e EKIN [16] agregaram ao seu sistema a análise de deslocamento de objetos no tempo, enquanto CHANG *et al.* [36], LAZARESCU *et al.* [35] e KIJAK *et al.* [50] também consideraram a movimentação de câmera para classificar com maior eficácia.

3.1.5 Outras Abordagens

Apesar de não serem tratadas neste trabalho, existem outras características bastante exploradas na literatura, tais como detecção de *replays*, rastreamento de objetos específicos via imagem de transmissão televisiva ou proveniente de sistemas que controlam câmeras e montam imagens 3D.

Detecção de *Replays*

Alguns algoritmos confiam no fato de que *replays* em câmera lenta são sempre construídos a partir da repetição de quadros de vídeo, o que faz com que o número de bits resultantes após a codificação fique abaixo do que em outras situações. Entretanto, ultimamente tem se tornado muito comum o uso de câmeras ultra-rápidas, o que inviabiliza algoritmos como os utilizados por KOBLA *et al.* [51] e KOBLA *et al.* [52], que generalizam o problema desta maneira.

Outra forma de tratar o problema foi apresentada por BABAGUCHI *et al.* [46], ASSFALG *et al.* [53] e ELDIB *et al.* [27], onde efeitos de edição são detectados e estes momentos são supostos como prováveis início e fim de um *replay*. Este método é eficiente, mas requer conhecimento prévio dos videografismos utilizados pela emissora ou campeonato em questão. Já PAN *et al.* [54] propuseram utilizar modelos HMM para identificar os *replays*, também utilizando, porém, a detecção de efeitos de edição para detectar os limites do segmento identificado.

Para superar as limitações relativas ao uso de videografismos, WANG *et al.* [55] realizam análises de contexto baseadas no tipo de cena e transição, enquanto WANG *et al.* [56] analisam a movimentação dos jogadores e da bola para detectar *replays*.

Rastreamento de Objetos a partir de Vídeos Transmitidos

O rastreamento de objetos a partir de transmissões é abordado em muitos trabalhos na literatura. Apesar de ser possível detectar em detalhes mínimos o que está ocorrendo na partida, este tipo de sistema é composto por técnicas de alto custo computacional.

Por exemplo, EKIN [16] detecta a presença da grande área e do juiz da partida enquanto YANG *et al.* [57] buscam as traves do gol. As marcas do campo, tais como linhas de meio-campo, lateral, grande e pequena área, costumam ser usadas como referência para localização dos participantes de uma partida de futebol, como proposto por GONG *et al.* [58]. Entretanto, como a observação do que ocorre em campo depende exclusivamente do que está sendo transmitido, em momentos nos quais as marcas não são visíveis, técnicas para estimação do movimento de câmera, como as descritas em 3.1.2, são utilizadas.

Outra abordagem amplamente utilizada é o *template matching*, onde imagens padrão dos objetos em cena são buscadas por todo o quadro de vídeo. Dentre eles, pode-se citar INTILLE e BOBICK [59] para futebol americano, SUDHIR *et al.* [25] para tênis, SEO *et al.* [28], MATSUI *et al.* [60], BEBIE e BIERI [61] e UTSUMI *et al.* [62] para o futebol.

INTILLE e BOBICK [63] e TOVINKERE e QIAN [64] utilizam estas técnicas para a determinação de eventos de interesse, enquanto REID e ZISSERMAN

[65], KIM *et al.* [66] e BRANCA *et al.* [67] montam cenas virtuais 3D com a localização exata dos objetos rastreados através das imagens transmitidas.

Rastreamento de Objetos a partir de Câmeras Controladas

Diferentemente dos trabalhos apresentados na seção anterior, aqui são vistos sistemas que utilizam câmeras próprias que não fazem parte das transmissões televisivas. Em outras palavras, essas câmeras são controladas pelo sistema que realiza o rastreamento dos objetos em campo, possibilitando, assim, o melhor posicionamento das câmeras para o sistema de rastreamento, e não para a transmissão do evento. Por outro lado, este tipo de sistema se limita ao rastreamento de jogadores, árbitros, linhas no campo, traves e bola, já que não se tem acesso às características cinemáticas que uma transmissão televisiva fornece. Por exemplo, não há como o sistema analisar que após determinado lance houve a exibição de *replay* e, dessa forma, inferir que algo interessante ocorreu na partida.

Durante o *Super Bowl XXXV* ocorrido em janeiro de 2001, foi utilizado um dos primeiros sistemas de rastreamento por câmeras controladas apresentado por KANADE [68]. Após, GUEZIEC [69] descreveu um sistema capaz de rastrear a bola com múltiplas câmeras para os canais ESPN. Enquanto isso, em PINGALI *et al.* [70] e PINGALI *et al.* [71] foi apresentado um sistema de rastreamento de objetos de partidas de tênis com o objetivo de criar *replays* virtuais, que nos dias de hoje são usados como forma de esclarecimento de jogadas onde o jogador se viu prejudicado pela marcação da arbitragem. TAKI *et al.* [72] também descreveram um sistema de rastreamento para coletar estatísticas dos objetos em campo para o caso de futebol. Recentemente, foi lançado o sistema comercial SportVU da STATS [73], que também rastreia todos os objetos utilizando múltiplas câmeras para o caso do futebol. Este sistema tem sido utilizado para, entre outras tarefas, coletar estatísticas dos jogadores no torneio Liga dos Campeões da Europa e nas partidas produzidas pela TV Globo no Brasil.

Apesar deste tipo de sistema se apresentar com eficiência para indexação de ocorrências e posterior sumarização de uma partida, em toda partida que o sistema vá ser utilizado é necessário montar uma estrutura de filmagem no local da partida, e é também necessário que um operador calibre o sistema a fim de definir quem é cada jogador e os limites do campo, o que exige um alto investimento.

3.2 Características de Vídeo

Em muitos trabalhos, o processamento de vídeo é a tarefa principal de onde é possível apontar diretamente se o que está ocorrendo é ou não um lance de interesse. No

entanto, nesta dissertação partimos da suposição que o vídeo não necessariamente precisa ser a principal fonte de características, mas sim, pode contribuir juntamente com características de outras naturezas, tal como o áudio.

Dessa forma, Seção 3.2.1 descreverá um método capaz de obter características baseadas na cor dominante de cada quadro de vídeo. A análise desta característica torna-se importante uma vez que ao detectar que o campo de jogo está presente, sabe-se que o quadro em questão contém conteúdo relacionado ao jogo, por exemplo.

Em seguida, a Seção 3.2.2 apresentará uma técnica que tem como objetivo obter características baseadas no movimento de câmera que podem ser utilizadas na identificação de quadros de vídeo como panorâmicos ou não-panorâmicos. Este tipo de informação pode ser interessante já que, segundo OWENS [19], câmeras panorâmicas são usualmente utilizadas somente em alguns momentos da transmissão esportiva. Por exemplo, imagens panorâmicas raramente são utilizadas em *replays*, enquanto *close-ups* costumam ocorrer com maior frequência quando o jogo não está em andamento.

3.2.1 Cor Dominante

Durante uma transmissão de uma partida de futebol, sabe-se que não somente se exibe uma visão geral do campo de jogo, já que em diversos momentos são utilizados *close-ups* e imagens da audiência, como ilustrado na Figura 3.1. Também é notório que campos de futebol são predominante verdes, excetuando condições raras e extremas como neve. Assim sendo, esta condição em que o campo é quase sempre verde será explorada no sistema que está sendo proposto neste trabalho.

Porém, o método de extração de características tem de ser robusto às diversas variações de cores do campo que podem ocorrer, tanto entre um estádio e outro quanto no mesmo estádio ao longo da partida. Isto é, a tonalidade do verde encontrado em um estádio muito provavelmente será diferente da encontrada em outro, e condições climáticas influenciam diretamente na tonalidade do verde de cada estádio. Além disso, no mesmo estádio, um jogo pode começar de tarde e terminar à noite, sofrendo uma intensa mudança em sua iluminação.

Espaço de Cores HSI

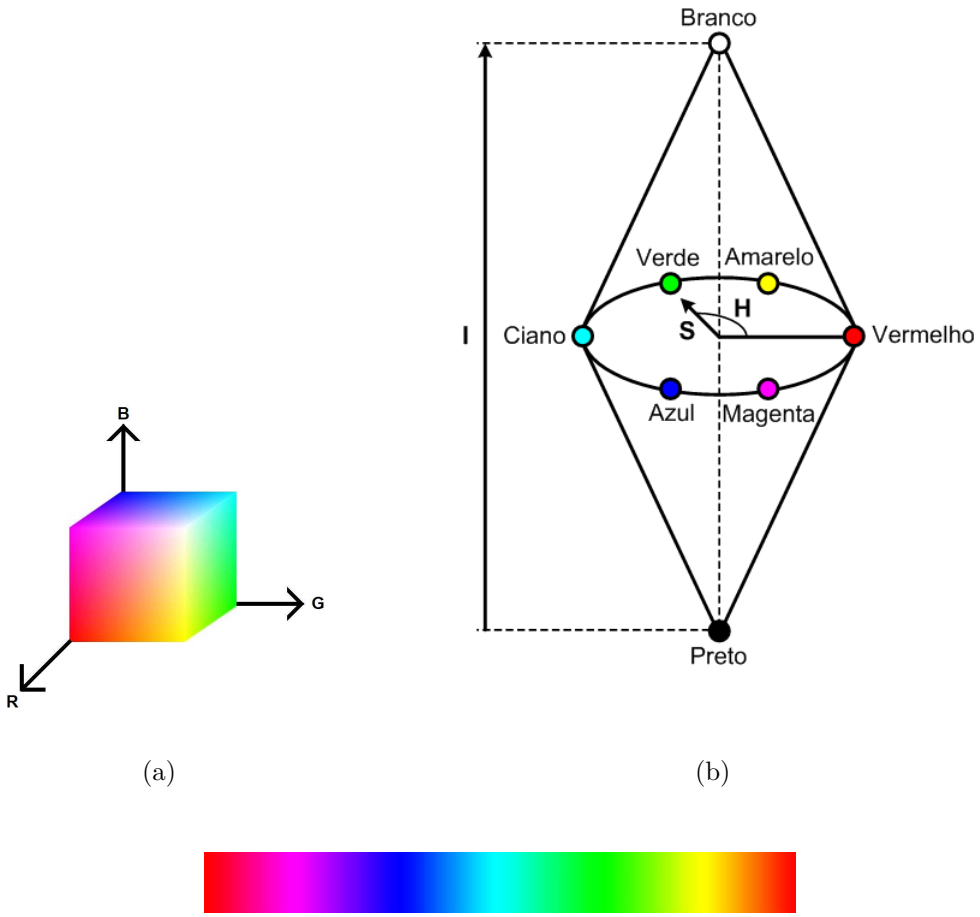
Por este motivo, é importante usar um espaço de cores capaz de discriminar bem as cores, que parecem diferentes para os humanos, e não discriminar as muito semelhantes. É notório (JAIN [74]) que o sistema RGB não possui esta característica. Um sistema de cores mais adequado para isto é o sistema HSI (matiz-saturação-intensidade).

Figura 3.1: Quadros comuns em uma transmissão de futebol para TV. Nota-se que quadros que incluem o campo de jogo contém uma parcela considerável do quadro na cor verde.

JAIN [74] afirma que a componente matiz H significa o que um leigo chama de cor, a cor pura que o olho realmente vê, tais como vermelho, verde, laranja e amarelo. Já a componente saturação S representa a pureza da cor, ou seja, o quanto a cor pura está diluída pelo branco. Por fim, a componente intensidade I representa o brilho, a luminosidade da cor. Pelo fato de ser mais próximo à percepção de cores existente na visão humana, este modelo de separação das componentes de cor atende melhor a algoritmos que precisam interpretar as cores, como o caso deste trabalho.

A componente matiz H varia de acordo com o ângulo do espaço de cores HSI representado na Figura 3.2. Pode-se afirmar que somente esta componente é suficiente para definir uma cor pura, enquanto o espaço RGB necessita dos valores das três componentes para esta definição. As exceções ocorrem nos casos das cores preta e branca que são atingidas neste modelo com valores muito baixos e muito altos, respectivamente, da componente intensidade I que é representado pelo eixo de simetria da figura. Da mesma forma, quando a saturação S, que é a amplitude do vetor indicado na figura, se aproxima do zero, o matiz H deixa de ter sentido,

o que resulta na prática em cores na escala de cinza. Estas regiões costumam ser denominadas como regiões de acromaticidade.



(a) (b)

(c)

Figura 3.2: Espaços de cor RGB (a) e HSI (b). Ilustração da faixa de valores do matiz (c) do espaço HSI, onde é fácil notar que uma larga faixa de matizes é atingida. Com isto, torna-se fácil com uma só componente discriminar o verde das outras cores, por exemplo.

No entanto, durante o processamento de vídeo não é comum decodificadores de vídeo entregarem os seus quadros no espaço HSI, o que obriga aos algoritmos implementarem etapas de conversões entre espaços de cores. Neste trabalho, o decodificador em questão disponibiliza o quadro no espaço RGB, portanto, serão utilizadas fórmulas de conversão citadas por MYLER e WEEKS [75] como equações eficientes e rápidas. Por serem muito complexas e estarem fora do escopo deste trabalho, não serão demonstradas as derivações necessárias para se chegar a essas equações, que, porém, podem ser encontradas em GONZALEZ e WOODS [76].

O primeiro passo para conversão de espaço RGB para o espaço HSI é normalizar

as componentes primárias RGB, como descrevem as Equações (3.1)-(3.3).

$$r = \frac{R}{R+G+B} \quad , \tag{3.1}$$

$$g = \frac{G}{R+G+B} \quad , \tag{3.2}$$

$$b = \frac{B}{R+G+B} \quad . \tag{3.3}$$

A partir disto, as componentes matiz H, saturação S e intensidade I podem, então, ser obtidas através de

$$A = \sqrt{\left(r-\frac{1}{3}\right)^2 + \left(b-\frac{1}{3}\right)^2 + \left(g-\frac{1}{3}\right)^2} \quad , \tag{3.4}$$

$$B = \frac{2}{3}(r-\frac{1}{3}) - \frac{1}{3}(b-\frac{1}{3}) - \frac{1}{3}(g-\frac{1}{3}) \quad , \tag{3.5}$$

$$\theta = \arccos\left(\frac{B}{A\sqrt{\frac{2}{3}}} - \frac{180}{\pi}\right) \quad , \tag{3.6}$$

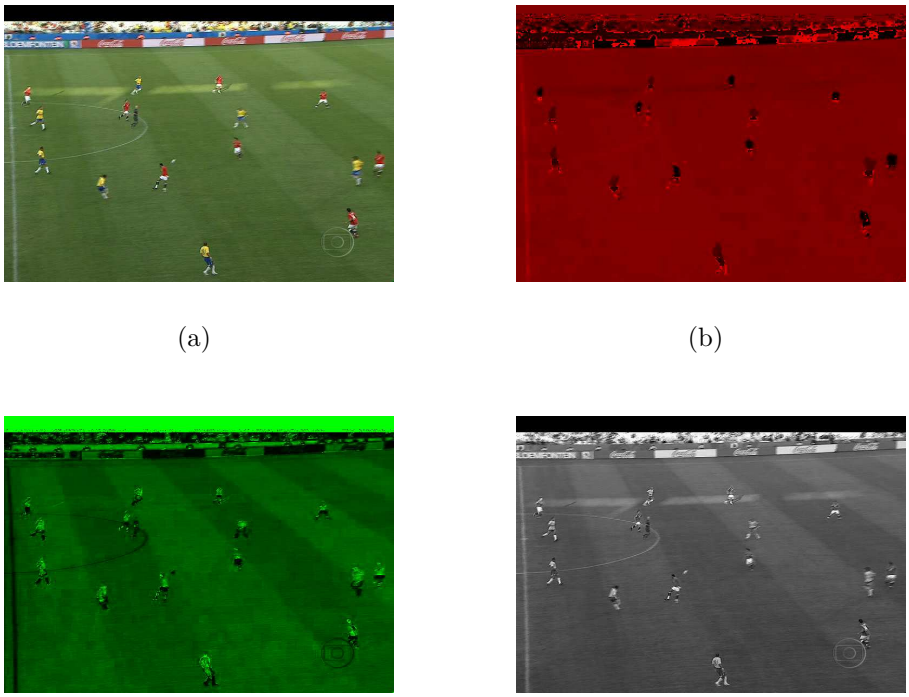
$$H = \begin{cases} \theta & \text{, se } g \geq b; \\ 360^\circ - \theta & \text{, caso contrário.} \end{cases} \quad , \tag{3.7}$$

$$S = 1 - 3\min(r,g,b) \quad , \tag{3.8}$$

$$I = \frac{R+G+B}{3} \quad . \tag{3.9}$$

Por conveniência o matiz H foi definido entre 0° e 360°, a saturação S entre 0 e 1, e a intensidade I entre 0 e 255.

Em princípio, sempre se pensa em utilizar na discriminação e análise das cores o histograma das três componentes HSI. Porém, é interessante estudar o comportamento de cada componente para o caso do futebol, para, então, entender a importância de cada uma delas. Mas para isso foi necessário desenvolver uma forma de visualizar cada uma delas. Assim, a componente H de cada *pixel* HSI foi quantizada para 8 bits e plotada na componente R do pixel RGB de mesma posição na imagem. A componente S do HSI teve o mesmo processo para a componente G do RGB e a componente I teve processo similar, só que sendo plotada em todas as componentes RGB. A Figura 3.3 exemplifica esta maneira de exibir as componentes HSI na imagem RGB. Em todos os casos quanto mais escuro o *pixel*, menor o valor da componente.



(a) (b)

(c) (d)

Figura 3.3: Imagem RGB original (a), imagens de componentes matiz (b), saturação (c) e intensidade (d). Note que as componentes saturação e intensidade não são necessárias para a simples diferenciação do verde para as outras cores.

Ao analisar a Figura 3.3 e considerando o que já foi discutido é possível reparar que a componente matiz H já é suficiente para determinar qual é a cor de um dado *pixel*, pois este tem valores diferentes para cores diferentes e valores próximos para tons diferentes da mesma cor. Isto se torna bem claro ao notar as imagens de exemplo, onde as componentes saturação S e intensidade I são capazes de diferenciar as faixas do gramado que são verdes mas em tonalidades diferentes, enquanto na componente H é praticamente impossível encontras estas faixas. Dessa forma, é vantajoso simplificar o cálculo de cor dominante para utilização somente da componente matiz.

Assim, pode-se pensar, então, que as componentes de saturação S e intensidade I poderiam ser desconsideradas, pois estão relacionadas ao brilho e quantidade de branco misturadas na cor. Entretanto, ao observar a Figura 3.2, pode-se notar que se o valor da saturação tiver um valor próximo de zero e a intensidade um valor próximo do zero ou do valor máximo, o matiz terá um valor indeterminado. Portanto, faz-se necessário o cálculo de saturação e intensidade para decidir quando o valor do matiz deve ou não ser considerado.

Estimação da Cor Dominante

Uma vez definido o espaço de cores HSI como o ideal para extração de informações relacionadas a cor para o caso do futebol, o próximo passo é estudar os cálculos necessários para obter uma única cor que representa um quadro de vídeo e que pode indicar se este quadro possui ou não um campo de futebol.

Usando como base o estudo realizado por EKIN [16], esta dissertação também aplica análises estatísticas em cada quadro de vídeo para determinar a sua cor dominante. Em resumo, a idéia basicamente consiste em calcular a média ao redor do pico do histograma de cada componente obtida da imagem como ilustrado na Figura 3.4.

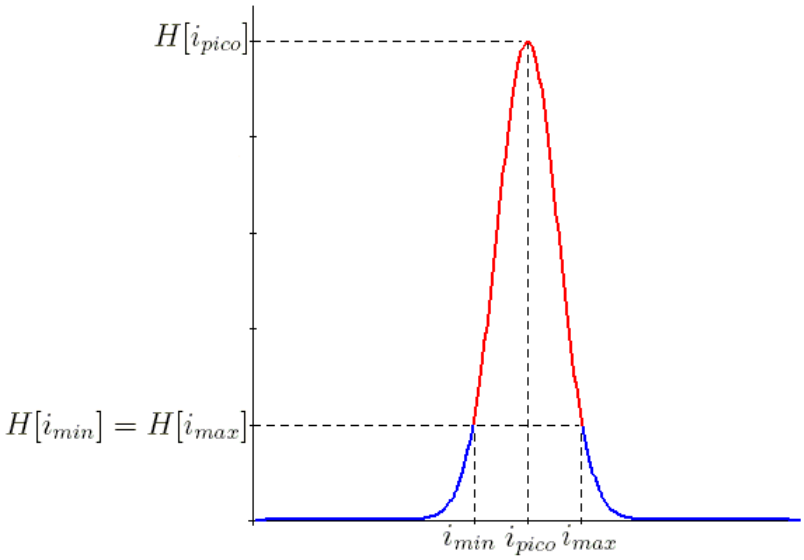


Figura 3.4: Histograma da componente matiz do espaço de cores HSI.

Porém, como já dito, as componentes saturação S e intensidade I não têm influência direta na classificação de um *pixel* como verde (para o caso do futebol), tornando os seus histogramas de pouca utilidade. Mas, também como já dito, S e I são necessários para definir quando o *pixel* pertence à região de acromaticidade, o que torna estas componentes úteis simplesmente para descartar este *pixel* do histograma da componente matiz. Em outras palavras, será calculado somente o histograma da componente matiz H, e as componentes saturação S e intensidade I serão usadas para não incluir determinados *pixels* em região de acromaticidade no histograma da matiz.

Após os histogramas serem computados, são aplicadas as Equações (3.10)-(3.16) com o intuito de encontrar o valor dominante da componente calculada definida por

ValorDominanteMatiz (ver Equação (3.16)), já que o pico do histograma representa o valor que mais ocorreu na imagem. Os valores dos índices i_{pico}, i_{min} e i_{max} úteis ao cálculo da média de cor dominante retirada do histograma devem obedecer as equações abaixo:

$$H[i_{min}] \geq K * H[i_{pico}] \quad , \tag{3.10}$$

$$H[i_{min} - 1] < K * H[i_{pico}] \quad , \tag{3.11}$$

$$H[i_{max}] \geq K * H[i_{pico}] \quad , \tag{3.12}$$

$$H[i_{max} + 1] < K * H[i_{pico}] \quad , \tag{3.13}$$

$$i_{min} \leq i_{pico} \quad , \tag{3.14}$$

$$i_{max} \geq i_{pico} \quad , \tag{3.15}$$

$$ValorDominanteMatiz = \frac{\sum_{i=i_{min}}^{i_{max}} H[i] * i}{\sum_{i=i_{min}}^{i_{max}} H[i]} \quad . \tag{3.16}$$

A constante K representa o percentual mínimo de ocorrência para i_{min} e i_{max} em relação ao i_{pico}, que neste trabalho foi fixado em 20%, ou seja, $K = 02$.

A Equação (3.16) explicita o cálculo que deve ser feito na região entre i_{min} e i_{max} do histograma para se obter $ValorDominanteMatiz$. Este tipo de abordagem faz com que o cálculo da média desconsidere os valores de matiz muito distantes do valor que ocorre mais frequentemente.

Os cálculos supra-citados podem ser aplicados diretamente no histograma da imagem. Entretanto, foi observado que em algumas imagens que continham pequenas inserções gráficas, um valor específico da componente ocorria mais vezes do que o valor referente ao campo de futebol, mesmo este participando de uma grande porção da imagem. Isto é justificado pelo fato de que inserções gráficas são geradas artificialmente, sendo possível que contenham o mesmo valor de componente em diversos *pixels* em pequenas áreas gerando um pico pronunciado no histograma. Por outro lado, o valor referente ao verde do campo de futebol sofre variações inerentes a imagens naturais, gerando menos picos pronunciados. O resultado disso é que apa-

recem dois picos pronunciados no histograma, sendo que o pico relativo à inserção gráfica não possui valores vizinhos com número de ocorrências satisfatórias.

Portanto, apesar de ocorrer em maior número, inserções podem não ser consideradas como dominantes se analisadas visualmente. Assim, foi aplicado um filtro de média móvel no histograma antes das equações serem calculadas. Isto fará com que picos espúrios percam importância em relação a picos que têm valores vizinhos com um número de ocorrências considerável, e, desta forma, valorizando as imagens naturais em detrimento das artificiais que não fazem parte do escopo deste trabalho.

Região do Campo na Imagem

Após estimar a cor dominante da imagem, é possível usar esta informação para obter outros dados, como o percentual de *pixels* de uma imagem que pertencem a valores próximos à cor dominante encontrada. Este resultado equivale aproximadamente ao percentual da imagem que pertence ao campo de futebol.

Para isso, deve-se então calcular a distância d_{matiz} de cada *pixel* para o valor dominante *ValorDominanteMatiz* da imagem. As Equações (3.17) e (3.18) indicam como é calculada a distância d_{matiz} para cada *pixel* j da imagem.

$$\Delta(j) = |\bar{H} - H_j| \quad , \tag{3.17}$$

$$d_{matiz}(j) = \begin{cases} \Delta(j) & , \text{se } \Delta(j) \leq 180°; \\ 360° - \Delta(j) & , \text{caso contrário.} \end{cases} \quad . \tag{3.18}$$

Uma vez calculadas as distâncias entre os *pixels* e a cor dominante, é possível definir um limiar para separar os *pixels* que são regiões do campo dos *pixels* que não são do campo. Neste trabalho, experimentalmente chegou-se ao limiar de 30°.

Resultados Preliminares

Ao fim da formulação do método de extração de características relacionadas à cor dominante do quadro de vídeo, é útil aplicá-lo a algumas imagens corriqueiras em transmissões televisivas de futebol com o objetivo de avaliá-lo.

O primeiro experimento, exibido na Figura 3.5, foi realizado em uma imagem que é quase totalmente coberta por regiões de campo. Na máscara binária gerada após a aplicação do limiar de decisão de campo pode-se ver que o algoritmo obteve um bom resultado ao eliminar os *pixels* que não são do campo e manter os que são. Além de *pixels* do campo, restaram alguns outros que estão em uma região que é caracterizada por possuir grande quantidade de detalhes, logo muitas cores e ocasionalmente *pixels* isolados que estejam próximos à cor verde.

Figura 3.5: Quadro de vídeo original e máscara binária para exemplo 1. A região do campo foi identificada com sucesso.

Posteriormente foi feita outra experiência com uma imagem muito parecida com a anterior, porém, além do campo, vê-se uma parcela maior do estádio, como visto na Figura 3.6. Com isso, como o número de *pixels* do campo é menor do que no caso anterior, poderia acontecer da cor dominante não ser tão próxima à cor do campo. Porém, ao observar a máscara binária resultante, nota-se que, também para este caso, o algoritmo foi capaz de identificar corretamente quais os *pixels* são do campo com sucesso. Da mesma maneira que na experiência anterior, repara-se que fora da região do campo, alguns *pixels* foram marcados como cor dominante. A explicação para tal é a mesma anterior, mas pode-se considerar também que é possível que alguns destes *pixels* estão na região de acromaticidade do espaço de cores HSI e, portanto, podem não ter sido bem definidos pelo algoritmo.

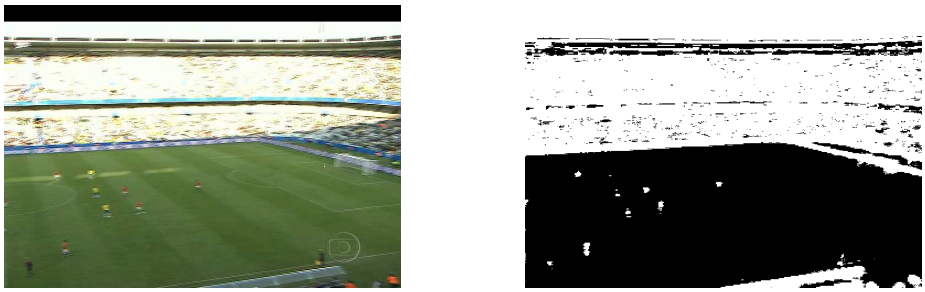


Figura 3.6: Quadro de vídeo original e máscara binária para exemplo 2. Ruídos fora da região de campo são encontrados, mas nem por isso deixa de ficar claro onde está o campo de jogo.

Na Figura 3.7, está o terceiro experimento que foi realizado com uma imagem que contém o campo, placas de publicidade, arquibancada e goleiro com uma camisa verde com um tom diferente do verde do campo. Como resultado, a máscara binária mostra que o campo mais uma vez foi identificado com sucesso, e, apesar de não ser o mesmo tom de verde, a camisa do goleiro também foi identificada, como esperado durante o desenvolvimento do algoritmo. Mais uma vez, devido às razões

já explicadas, há ruídos na região da arquibancada.

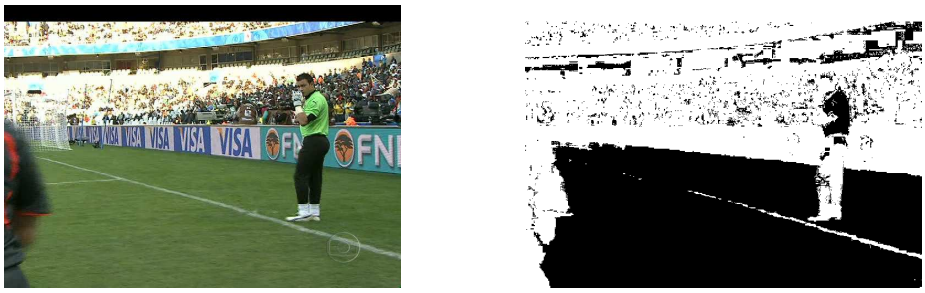


Figura 3.7: Quadro de vídeo original e máscara binária para exemplo 3. Nota-se que o limiar de 30° é suficiente para incluir a camisa com tonalidade de verde diferente do campo, e descartar as placas de publicidade atrás do goleiro na cor ciano, mesmo esta sendo bem próxima ao verde em valores de matiz.

No experimento seguinte, apresentado na Figura 3.8, foi utilizada uma imagem onde a proporção de *pixels* de campo pelo total da imagem é menor do que nos experimentos anteriores e há objetos dominando maiores porções da imagem, como a camisa do jogador em amarelo. Isto poderia nos levar a crer que a componente matiz da cor dominante calculada esteja distante do matiz da cor do campo. Entretanto, pelo resultado exposto na máscara binária, vê-se que o algoritmo novamente conseguiu identificar os *pixels* do campo satisfatoriamente. Novamente, o ruído nas regiões pertencentes à arquibancada e à estrutura do estádio ocorreram, porém, desta vez, com maior peso. Neste caso, regiões de cinza foram identificadas como cor do campo, o que pode ser explicado pelo fato do cinza não possuir saturação, tornando possível que o matiz assuma qualquer valor, podendo então ser confundido com o verde.

Figura 3.8: Quadro de vídeo original e máscara binária para exemplo 4. O algoritmo mostra-se capaz de determinar quais os *pixels* de uma imagem são do campo, mesmo quando o percentual destes na imagem é reduzido.

Por fim, no último experimento, exibido na Figura 3.9, foi utilizada uma imagem que não contém sequer um *pixel* de campo. Assim, o algoritmo atribuiu à cor

dominante um valor de matiz que deve ser próxima do amarelo e distante do verde. Com isso, a máscara binária resultante identifica as camisas amarelas, e parte da pele das pessoas que estão na imagem, o que está de acordo com o esperado para este algoritmo, mas não se aplica a este método de identificação do campo de futebol. Porém, este caso possivelmente não deverá criar problemas para o sistema, uma vez que o percentual do campo pertencente à cor dominante será aliado ao valor absoluto da cor dominante no classificador que será apresentado no Capítulo 4. Assim, o classificador deverá entender que ao identificar cores dominantes distantes do verde, a imagem em questão deverá ser desconsiderada.

Figura 3.9: Quadro de vídeo original e máscara binária para exemplo 5. Máscara binária foi corretamente formada mesmo quando a cor dominante não foi o verde.

Com isso, o método utilizado agrega duas características às já existentes, sendo a primeira o valor absoluto da cor dominante do quadro, e a segunda, o percentual dos *pixels* do quadro que estão próximos à cor dominante encontrada. Estas características não são conclusivas para a determinação de momentos de interesse. Porém, estas podem ajudar neste processo: no instante, por exemplo, em que a cor dominante do quadro não é verde muito provavelmente neste instante da transmissão não está ocorrendo um momento de interesse.

3.2.2 Movimento de Câmera e Quadros Panorâmicos

As Figuras 3.10 e 3.11 mostram que equipes de produção de partidas de futebol em transmissões de TV sempre utilizam mais de uma posição da câmera, tais como as que estão no nível do campo a uma pequena distância, as que estão um pouco acima do nível do campo a uma média distância e as panorâmicas que ficam no nível mais alto que a infra-estrutura do estádio permite e, por consequência, distantes do campo de jogo.

Além disso, KOKARAM *et al.* [77] definem o futebol como um esporte MVS (*Multiple View Semantics*), onde somente uma posição de câmera não é capaz de cobrir toda ação do jogo, enquanto em esportes DSV (*Dominant Semantic View*), tais como tênis, uma posição de câmera é suficiente.

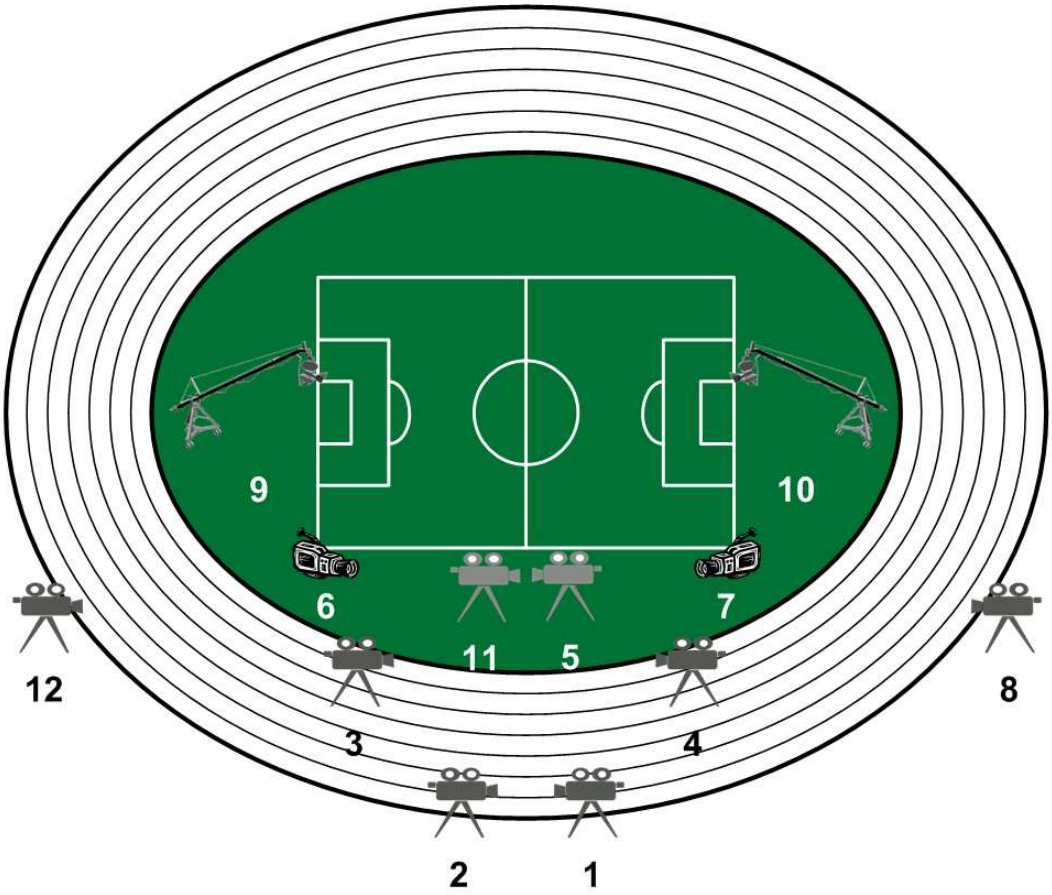


Figura 3.10: Exemplo de disposição das câmeras em uma partida de futebol. Câmeras 1, 2, 8 e 12 são usadas para cenas panorâmicas, enquanto 3, 4, 9 e 10 para cenas de média distância e da audiência, e, por fim, 5, 6, 7 e 11 capturam imagens no nível do campo como *close-ups*. Ilustração baseada em OWENS [19].

Aliando estes dois fatos ao já descrito pela literatura na Seção 3.1, chega-se à conclusão que a partir da análise de movimentação de câmera e o tipo de tomada de câmera é possível obter informações relevantes sobre as características cinemáticas aplicadas pelas equipes de produção de transmissões esportivas.

Entretanto, vale ressaltar que esta característica está diretamente associada a forma como as equipes de produção realizam a transmissão da partida, tornando esta característica sensível a diferenças na produção de equipes de TV distintas. Apesar desta sensibilidade, nos dias de hoje as equipes de produção costumam trabalhar de forma bastante similar, essencialmente no uso da câmera panorâmica em momentos em que ações de perigo ocorrem ao vivo, o que é suficiente para tornar o uso desta característica interessante para o sistema.

Estimação do Movimento de Câmera

O algoritmo para estimação de movimento de câmera consiste em inferir o movimento global da imagem e seus objetos. Caso, por exemplo, o algoritmo aponte

Figura 3.11: Imagens de câmera panorâmica, *close-up* e da audiência são comuns em transmissões de futebol.

movimento global dos objetos para a direita, supõe-se que a câmera está se movimentando para a esquerda. Para isso, é comum utilizar cálculos de correlação entre quadros de vídeo subsequentes, que é uma forma mais ágil de realizar esta estimação comparada a algoritmos baseados em rastreamento de objetos, que são mais pesados computacionalmente.

Ao seguir esta linha, é formulada em PEARSON [78] a correlação por fase da seguinte maneira: ao considerar uma imagem I_1 que sofre um movimento de translação representado por um vetor (v_x, v_y), pode-se estimar a imagem I_2 subsequente como

$$I_2(x, y) = I_1(x - v_x, y - v_y). \tag{3.19}$$

Ao aplicar a transformada de Fourier em ambos os lados, tem-se

$$F_2(m, n) = F_1(m, n)e^{-\pi j(mv_x + nv_y)}, \tag{3.20}$$

onde F representa a transformada de Fourier, enquanto m e n as frequências do espectro resultante da transformada (DINIZ *et al.* [24]).

A partir disso, pode-se chegar, então, à transformada de Fourier da correlação cruzada através de

$$F_2(m, n) = F_1F_2^* = F_1F_1^*e^{2\pi j(mv_x + nv_y)}. \tag{3.21}$$

E, ao dividir a Equação (3.21) por $F_1F_1^*$ obtem-se

$$C(x, y) = \delta(x - v_x, y - v_y), \tag{3.22}$$

que será uma função de δ que representa o deslocamento entre as imagens I_1 e I_2. Entretanto, para o caso prático de transmissões esportivas não ocorre um movimento de translação puro entre duas imagens, pois usualmente há outros objetos se movendo simultaneamente, tais como jogadores, bola e audiência. Com isso, pode-se generalizar o cálculo de correlação cruzada para

$$C(x,y) = F^{-1}\left[\frac{F_1 F_2^*}{|F_1 F_2^*|}\right], \tag{3.23}$$

onde F_1 e F_2 são as transformadas de Fourier de quadros subsequentes enquanto F^{-1} é a transformada inversa de Fourier.

Este cálculo é interessante por somente utilizar FFTs e operações multiplicativas, tornando-o mais rápido computacionalmente que as correlações temporais. A Equação (3.23) define o mapa de correlação 3D $C(x,y)$ que foi exemplificado na Figura 3.12, onde o maior pico representará o movimento dominante entre os quadros.

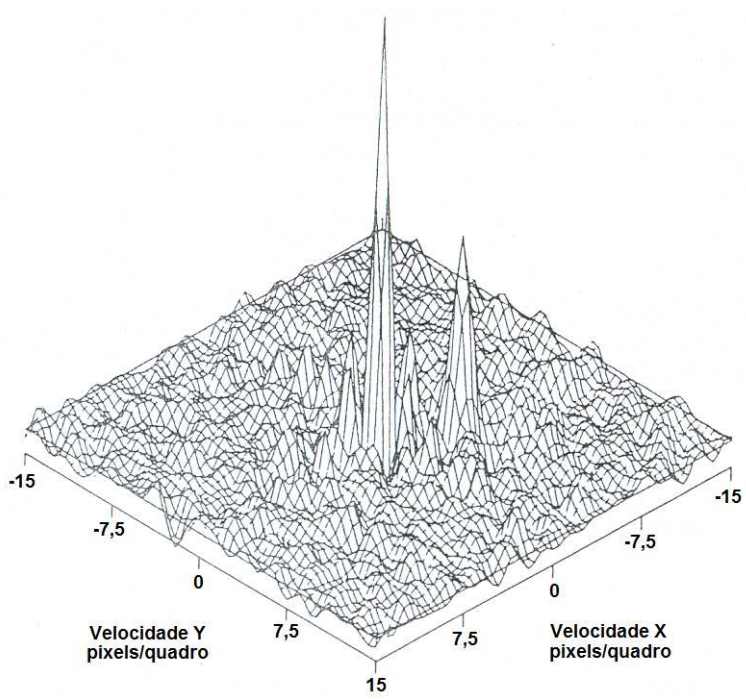
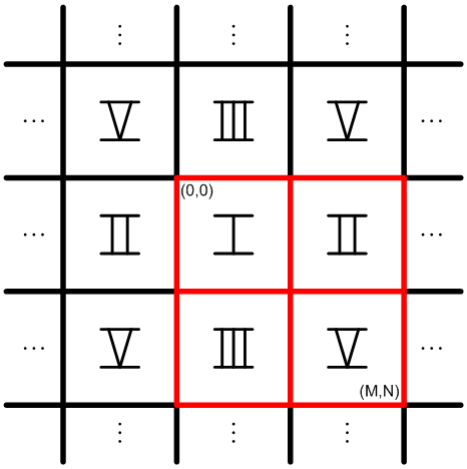


Figura 3.12: Mapa 3D de correlação por fase entre duas imagens. Fonte: PEARSON [78].

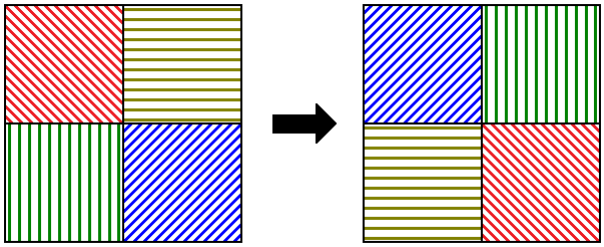
Além disso, pelo fato de outros objetos estarem se movendo em diferentes direções entre os quadros, vê-se que é comum aparecerem outros picos que representem estes movimentos secundários. Porém, quanto menores forem as áreas cobertas por objetos, menores serão os picos resultantes. Isto nos leva a concluir que a amplitude do pico pode ser utilizada como uma medida de confiabilidade da estimativa do movimento dominante entre os quadros.

A Figura 3.13a exibe em destaque uma visão 2D do mapa $C(x,y)$ resultante da Equação (3.23), de onde, a partir do ponto $(0,0)$, é possível quantificar, em número de *pixels* horizontais e verticais, o movimento global. Ao utilizar DFTs, um movimento para esquerda e para cima provoca picos próximos às bordas do mapa devido às repetições do espectro que aparecem em preto na figura.

Portanto, para um melhor entendimento do mapa e manuseio de seus dados, foi



(a)



(b)

Figura 3.13: Mapa 2D de correlação por fase (a) representado em vermelho dividido em quadrantes e suas repetições provenientes da FFT em preto. Inversão dos quadrantes (b). Mapa 3D de correlação por fase com quadrantes invertidos (c).

aplicada a inversão dos quadrantes ilustrada na Figura 3.13b, fazendo com o que a origem $(0, 0)$ ficasse, assim, sempre no centro do mapa, como mostra a Figura 3.12. Em geral, como movimentos de um quadro para outro costumam ser pequenos, os picos dominantes se concentrarão no centro do mapa.

Transformação de Coordenadas

Definido o mapa 3D de correlação que indicará para onde a câmera está se movendo, é necessário extrair parâmetros que representem este movimento. Em um primeiro momento, imagina-se que a melhor maneira é utilizar o número de *pixels* de distância, tanto na horizontal quanto na vertical, entre o maior pico e o centro do mapa.

No entanto, ao aplicar uma transformação do sistema de coordenadas cartesianas para a polar, indicada pelas Equações (3.24) e (3.25), são obtidas informações mais significantes para a estimação de movimento da câmera:

$$\Delta = \sqrt{x^2 + y^2} \quad , \tag{3.24}$$

$$\theta = \text{arctg}\left(\frac{y}{x}\right) \quad . \tag{3.25}$$

Isto significa que, dessa forma, a magnitude Δ do vetor oriundo da origem para a base do maior pico do mapa 3D será descrita pelo tamanho do movimento entre os quadros, enquanto o ângulo θ entre este vetor e uma referência indicará a direção e sentido do movimento entre os quadros. Além disso, pode-se definir a amplitude ρ do pico 3D, que representa o quão bem-definido é este pico, ou seja, quanto maior o valor deste pico, maior a certeza que este representa o movimento global entre os quadros.

Análise das Características

Após a criação dos parâmetros magnitude Δ, direção θ e confiabilidade ρ para explorar o mapa resultante do movimento ocorrido na imagem de um quadro para o seu subsequente, deve-se analisar como estes parâmetros podem contribuir para que características cinemáticas da produção do evento esportivo sejam identificadas. Como os parâmetros criados são relacionados ao movimento dominante da imagem, é interessante analisar a variação destes parâmetros ao longo do tempo.

Além de informações óbvias, tais como para onde a câmera está se movendo e em que velocidade, supõe-se que a variação dos parâmetros Δ, θ e ρ ao longo do tempo pode indicar outros tipos de características, como qual câmera está sendo utilizada naquele momento. Isto acontece porque imagina-se que, por exemplo, em imagens panorâmicas, por conterem objetos menores na imagem, haja movimentos com velocidade menor do que em *close-ups*, onde os objetos estão mais próximos e consequentemente gerando movimentos mais bruscos. Portanto, a fim de obter conclusões sobre como estes parâmetros se comportam nestes casos, foram extraídos os parâmetros Δ, θ e ρ vistos na Figura 3.14 para um trecho onde primeiro ocorre uma cena de *close-up* em um jogador (cena 1), seguida por uma cena panorâmica do campo de jogo (cena 2), uma cena da audiência presente no estádio (cena 3), e, finalmente, por mais uma cena panorâmica do campo de jogo (cena 4).

Na Figura 3.14 é fácil notar que nos trechos panorâmicos, os parâmetros Δ e θ variam consideravelmente menos do que em trechos de *close-up* e da audiência. Isto é explicado pelo fato de que em *close-ups*, os objetos em cena são maiores, e se deslocam em mais direções e maiores deslocamentos. Por exemplo, em um *close-up* de um jogador, cada braço pode se mover em direções e velocidades diferentes, e além disso, o braço pode mudar de direção e velocidade em poucos quadros, o que

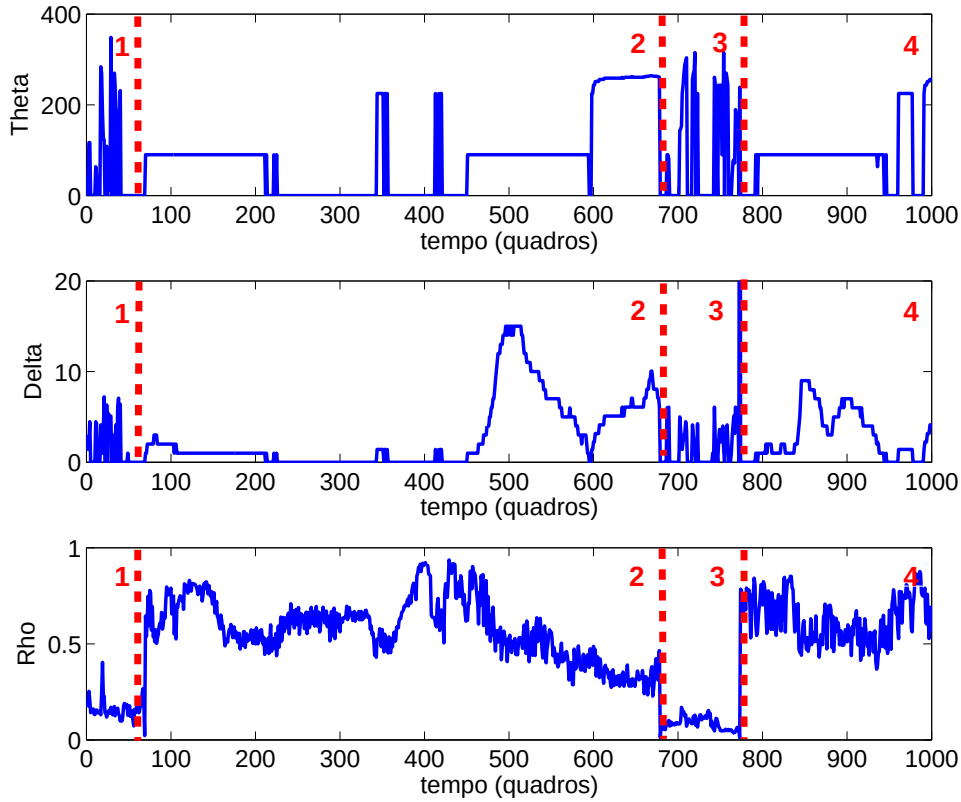


Figura 3.14: Parâmetros θ, Δ e ρ ao longo do tempo para vídeo contendo sequências de imagens não-panorâmicas, 1 e 3, e panorâmicas, 2 e 4. Em cenas panorâmicas, θ e Δ se mantêm estáveis enquanto em não-panorâmicas, variam bastante.

provocará mudanças drásticas nos valores de Δ e θ. Já em tomadas panorâmicas, os objetos são bem menores e seus movimentos são pequenos em proporção à área de jogo da imagem, e, assim, os *pixels* do campo e estádio se moverão na mesma direção e velocidade, ocasionando valores de Δ e θ mais constantes ao longo do tempo.

Também, nota-se na Figura 3.14 que a amplitude ρ do pico do mapa 3D apresenta valores maiores nas cenas panorâmicas. Pelo mesmo motivo, ρ em cenas panorâmicas tende a ser maior, pois o movimento global é melhor definido já que quase toda a imagem se move nas mesmas direção e velocidade. Por outro lado, utilizando o mesmo exemplo de *close-up* em um jogador, cada parte do braço em movimento gerará um pico em determinada posição do mapa 3D, o que faz com que não haja um pico que caracterize o movimento dominante da imagem. Assim, as amplitudes desses picos serão menores do que no caso do pico dominante em cenas panorâmicas.

Apesar desta análise indicar quando uma cena é panorâmica, em geral ela não é capaz de diferenciar cenas de *close-ups* de cenas de audiência. A explicação é similar à do caso do *close-up*, ou seja, a imagem é quase totalmente preenchida com muitos objetos se movendo cada um para uma direção e em velocidades diferentes.

Portanto, isso nos leva a crer que a estabilidade de Δ e θ, e o valor de ρ podem apontar uma imagem como panorâmica ou não-panorâmica.

3.3 Características Pós-Processadas

A seção anterior descreveu como os parâmetros Δ, θ e ρ podem contribuir para detecção de quadros panorâmicos a partir de uma análise da variabilidade destes parâmetros ao longo do tempo.

Uma maneira de medir essa variabilidade é calcular as variâncias nos sinais de Δ e θ. Para isso, aplica-se uma janela retangular de comprimento N que é deslocada de amostra em amostra onde a variância será calculada.

Experimentalmente, adotou-se o comprimento de janela N valendo 15 quadros de vídeo, que para vídeos NTSC corresponderá a cerca de meio segundo. Este valor é razoável, pois entende-se que dificilmente haverá mais de um tipo de cena neste período de tempo, e no máximo haverá uma troca entre dois tipos. Mesmo se houver uma troca de cena neste período, os sinais pós-processados não serão afetados consideravelmente, pois amostras da cena anterior a atual serão utilizadas por no máximo meio segundo. Vale notar que este tamanho de janela não pode ser muito pequeno, pois isto tornaria possível valores de variâncias bem distintos em quadros de vídeo próximos.

Na Figura 3.15 são apresentados os sinais de variância de θ e Δ aplicados ao mesmo trecho de vídeo da Figura 3.14. Nota-se que a intensidade do sinal de variância de θ é maior nos segmentos não-panorâmicos do que nos panorâmicos. Entretanto, também há instantes em que as intensidades de segmentos panorâmicos são similares às encontradas nos não-panorâmicos, o que pode causar falsos positivos na determinação do tipo de cena. Já no sinal de variância de Δ há poucos instantes de intensidade alta fora dos segmentos não-panorâmicos. Porém, há muitos instantes de baixa intensidade em segmentos não-panorâmicos, o que pode ocasionar falsos negativos.

Assim, pode-se usar as características Δ, θ e ρ, que contêm informação relacionada ao movimento global da câmera, e as variâncias de Δ e θ, que aliadas aos anteriores indicam quando um quadro é panorâmico, para alimentar um classificador tipo *AdaBoost*. Esta solução será abordada no Capítulo 4. Apesar de não indicar quando um gol ocorre, estas características podem ser úteis, por exemplo, em momentos de indecisão. Por exemplo, lances que estão acontecendo ao vivo costumam utilizar câmeras panorâmicas, e assim, muito provavelmente no momento do gol a imagem será panorâmica.

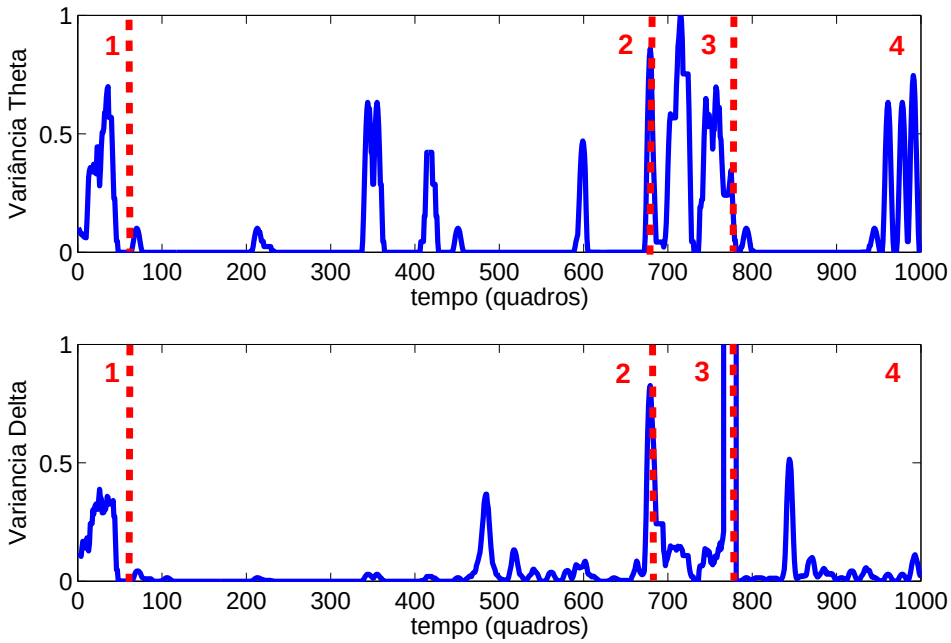


Figura 3.15: Variância θ e variância Δ ao longo do tempo para mesma sequência de vídeo.

3.4 Conclusões

Como o Capítulo 2, este capítulo estudou e desenvolveu características de vídeo que podem contribuir para o classificador de momentos de interesse em transmissões televisivas de futebol.

A Seção 3.1 revisou trabalhos que utilizam o vídeo de diversas maneiras diferentes para o tipo de sistema proposto por este trabalho. Foi visto, que muitos deles, assim como este, se aproveitam do fato de todas equipes de produção de TV fazerem uso das mesmas regras cinemáticas.

Depois disso, aa Seção 3.2.1 propôs um método baseado em análises estatísticas do espaço de cores HSI que aponta qual a cor dominante do quadro de vídeo e qual o seu percentual de *pixels* que estão próximos a esta cor dominante.

Já a Seção 3.2.2 estudou o movimento da câmera extraindo parâmetros que indicam para onde e em que velocidade a câmera está se movendo. Além disso, foi aplicado o operador de variância em dois desses parâmetros com o intuito de ser possível inferir quando a cena é panorâmica.

Com isto, as sete características geradas neste capítulo serão agregadas às características relativas ao áudio no classificador que será discutido no próximo capítulo. Assim como para o áudio, a real importância de cada característica para o sistema só será determinada após o treinamento do classificador.

Capítulo 4

Metodologia de Classificação

Os Capítulos 2 e 3 estudaram e propuseram características de áudio e vídeo que podem ser úteis na identificação de momentos de interesse em transmissões televisivas de futebol. Por este motivo, o objetivo principal deste capítulo é elaborar uma forma de agregar estas características utilizando um classificador com o intuito de se obter um sistema de identificação eficiente de melhores momentos.

Assim, na Seção 4.1 será revisto como outros trabalhos da literatura realizam a classificação de eventos para transmissões esportivas. Depois disto, a Seção 4.2 descreve o classificador *AdaBoost* e como ele será aplicado para o caso do sistema de melhores momentos proposto nesta dissertação. Em seguida, a Seção 4.3 discute a importância e a montagem da base de dados necessária ao treinamento do classificador. Posteriormente, na Seção 4.4 será apresentada a metodologia de validação do classificador. A Seção 4.5 descreve um roteiro de experimentação interessante para destacar a relevância das características de áudio e vídeo se usadas separadamente ou em conjunto no sistema proposto.

4.1 Revisão Bibliográfica

A princípio, a forma mais intuitiva de combinar características de naturezas diferentes é analisar a importância de cada uma delas na transmissão de determinado esporte para, posteriormente, definir que papel esta característica exercerá na classificação de eventos de maior nível semântico, tais como os melhores momentos de uma partida. Por isso, é muito comum em sistemas deste tipo que as regras de classificação sejam construídas a partir de análises da participação das diferentes características extraídas das partidas.

Isto é feito em TOVINKERE e QIAN [64] que aplicam uma sequência de regras heurísticas nos resultados do rastreamento dos objetos no campo de futebol para classificar as ações de cada jogador, assim como eventos ocorridos durante a partida. Em outro trabalho, COLDEFY e BOUTHEMY [17] uniram características de áudio

e vídeo com regras simples para determinar momentos de interesse em uma partida de futebol. Mais à frente, LAO *et al.* [79] realizam em primeiro lugar a detecção de batidas na bola, para em seguida aplicar regras de classificação para a identificação de eventos no caso do tênis, tais como *aces* (ponto marcado diretamente do saque), primeiro e segundo serviço, retorno para ponto e *rallies* (jogadas onde ocorrem várias trocas de bola entre os tenistas até a definição do ponto).

Dentre os classificadores encontrados na literatura, os mais populares são os baseados em HMMs (*Hidden Markov Models*). Por exemplo, XIONG *et al.* [80] e KIM *et al.* [81] utilizam HMMs para classificar os eventos de áudio em partidas de beisebol, golfe e futebol, enquanto LI e SEZAN [37], DENMAN *et al.* [82], DAHYOT *et al.* [83], DAHYOT *et al.* [6] e XIE *et al.* [38] modelam a evolução temporal de características de vídeo para classificação do tipo de cenas para diversos esportes.

Por ter a capacidade de tratar conteúdos de significado semântico, as HMMs também são muito utilizadas para modelar uma sucessão de eventos que indiquem algo de interessante, como fazem KIJAK *et al.* [50], ROZENN *et al.* [84] e REA *et al.* [85]. Além disso, LEONARDI *et al.* [14] e HUANG *et al.* [86] reúnem características de áudio e vídeo para elaborar um classificador baseado em HMMs para identificar os melhores momentos de uma partida de futebol e tênis.

Também é comum o uso de *Bayesian Network* (BN) e *Dynamic Bayesian Network* (DBN) para combinar os diversos aspectos de uma transmissão esportiva de TV. Por exemplo, VOJKAN *et al.* [9] e WANG *et al.* [87] baseiam-se em DBN para combinar características oriundas da análise de áudio, vídeo e texto para encontrar os melhores momentos em transmissões de Fórmula 1. Da mesma forma, HUANG *et al.* [88] separam em grupos os eventos gols, penalidades, cartões e escanteios ocorridos durante uma partida de futebol aplicando análises semânticas baseadas em BN e DBN.

Já OKUMA [89] e OKUMA *et al.* [90] construíram o *Boosted Particle Filter* (BPF) que é uma derivação do classificador *Adaptive Boosting* (*AdaBoost*) para rastrear automaticamente a trajetória dos personagens envolvidos em uma partida de *Hockey* no gelo. Após isto, MING *et al.* [91] propuseram uma variação do *AdaBoost* para combinar características de baixo-nível com o intuito de classificar cenas de transmissões esportivas como gelo, grama, neve e pisos artificiais.

Apesar de exigir estágios de treinamento com uma vasta base de dados, o uso de classificadores é considerado vantajoso, uma vez que após o treinamento, o classificador será capaz de indicar a importância de cada característica inserida durante o treinamento. A desvantagem dos classificadores para sistemas de identificação de melhores momentos é que em seu estágio de treinamento, a base de dados utilizada, em teoria, deve conter todos os exemplos possíveis de eventos, estádios, narradores que podem ocorrer durante uma partida, o que é uma tarefa muito laboriosa devido

à diversidade destes fatores encontrada no caso do futebol.

4.2 Classificador *AdaBoost*

De acordo com DUDA *et al.* [92], sistemas contendo classificadores são constituídos em cinco fases: sensoriamento, segmentação, extração de características, classificação e pós-processamento. O sensoriamento é onde a natureza que será observada é capturada e convertida em dados. No caso deste trabalho o vídeo e o áudio serão capturados durante a transmissão da partida de futebol. A segmentação é a fase onde o conteúdo obtido na fase de sensoriamento é separado de acordo com o exigido para o funcionamento do classificador, o que no caso deste trabalho não se aplica já que todo conteúdo capturado será utilizado durante a classificação da partida.

A próxima fase é a de extração das características que serão úteis ao classificador, que no caso deste trabalho estão de acordo com o já apresentado nos Capítulos 2 e 3, e que foram resumidas na Tabela 4.1. A fase de classificação é onde estas características são avaliadas com o objetivo de indicar a que classe pertence cada amostra. No caso deste trabalho será feita a classificação como bom momento ou lance normal. Por fim, a fase de pós-processamento é destinada a um eventual cálculo final com o intuito de melhorar os resultados.

Tabela 4.1: Características de áudio e vídeo que serão utilizadas pelo classificador neste trabalho.

Áudio		Vídeo	
Pitch	Energia	Cor	Movimento de Câmera
$p(t)$ (Seção 2.2.1)	$e_{st}(t)$ (Seção 2.2.2)	$h(t)$ (Seção 3.1.1)	$\Delta(t)$ (Seção 3.2.2)
$\mu_m^p(t)$ (Seção 2.2.3)	$\mu_m^{st}(t)$ (Seção 2.2.3)	$r(t)$ (Seção 3.1.1)	$\theta(t)$ (Seção 3.2.2)
$\mu_s^p(t)$ (Seção 2.2.3)	$\mu_s^{st}(t)$ (Seção 2.2.3)		$\rho(t)$ (Seção 3.2.2)
	$e_{cs}(t)$ (Seção 2.2.2)		$var(\Delta)(t)$ (Seção 3.2.2)
	$\mu_m^{cs}(t)$ (Seção 2.2.3)		$var(\theta)(t)$ (Seção 3.2.2)
	$\mu_s^{cs}(t)$ (Seção 2.2.3)		
	$e_{mcs}(t)$ (Seção 2.2.2)		
	$\mu_m^{mcs}(t)$ (Seção 2.2.3)		
	$\mu_s^{mcs}(t)$ (Seção 2.2.3)		

Dentre os classificadores existentes na literatura, os baseados no *Adaptive Boos-*

ting são interessantes quando o conjunto de classificadores e/ou características que o compõem são considerados como fracos. De acordo com SCHAPIRE [93] e FREUND [94], a idéia principal do *AdaBoost* é reunir características que não são determinantes para classificar um evento, mas que em conjunto com outras características fracas, constituam um classificador forte. KEARNS e VALIANT [95] consideram um classificador fraco o que atinge uma probabilidade de acerto em torno de 50%, enquanto um classificador forte atinge uma probabilidade próxima aos 100%. Uma descrição breve do classificador *AdaBoost* será apresentada nesta seção, mas sua formulação completa pode ser encontrada em SCHAPIRE e FREUND [96].

Também de acordo com DUDA *et al.* [92], a construção de um classificador, dentre eles o *AdaBoost*, é dividida em cinco etapas: coleta de dados, seleção de características, seleção do modelo, treinamento e validação. A primeira etapa consiste na montagem da base de dados para o posterior treinamento do classificador, que será vista com detalhes na Seção 4.3. A segunda etapa é onde as características que farão parte do classificador são selecionadas, pois é necessário ter conhecimento de quais delas realmente contribuirão para a classificação. Esta etapa no caso do *AdaBoost* é em teoria desnecessária, pois ele suporta tantas características quanto desejado e a falta de significado de determinada característica para o classificador não interfere no treinamento já que pesos irrisórios são associados a esta característica. Entretanto, é importante notar que, se entregamos ao classificador *AdaBoost* as características mais relevantes possíveis, 'facilitamos' o seu trabalho, o que pode diminuir a necessidade de treinamento.

A etapa de seleção do modelo indica como o classificador deve atuar, o que ele deve receber na entrada e o que deve indicar em sua saída, o que no caso deste trabalho é a entrada de áudio e vídeo oriundos de uma transmissão de TV de futebol e a saída dos melhores momentos da respectiva transmissão.

Em seguida, na etapa de treinamento, a base de dados é utilizada para ajustar os pesos de cada característica do classificador. Finalmente, a etapa de validação consiste em utilizar um conjunto de dados diferente do utilizado no treinamento para validar os resultados do classificador. Neste trabalho, foi utilizada a técnica da validação cruzada para tal tarefa, que será descrita mais adiante na Seção 4.4.

Para o *AdaBoost*, todas as etapas descritas podem ser resumidas na matriz da Equação (4.1) a seguir. As etapas de coleta de dados e extração de características são representadas em cada linha da matriz, menos a última. Isto é, cada linha da matriz possui todos os valores de uma das características $c(t)$ vista na Tabela 4.1 para cada amostra contida nos dados coletados. A etapa de seleção de modelo é representada tanto pelas linhas de característica já mencionadas, que indicam a entrada do classificador, quanto pela última linha onde a saída do classificador é tida pelos rótulos $r(t)$ de cada amostra, que indicam como cada amostra deve ser

vista pelo classificador, como um "bom momento" ou "lance normal" no caso deste trabalho.

Por fim, as etapas de treinamento e validação são representadas pela matriz completa, com a diferença de que os dados utilizados durante o treinamento não devem ser utilizados na validação. Outra diferença é que na etapa de validação a matriz não contém a última linha de rótulos, já que esta linha será o resultado do classificador.

$$\begin{array}{l} \\ \textit{Característica } 1 \\ \textit{Característica } 2 \\ \textit{Característica } 3 \\ ... \\ \textit{Característica } N \\ \textit{Rótulo} \end{array} \begin{array}{c} \begin{array}{cccc} \textit{Amostra } 1 & \textit{Amostra } 2 & ... & \textit{Amostra } M \end{array} \\ \begin{pmatrix} c_{11}(t) & c_{12}(t) & ... & c_{1M}(t) \\ c_{21}(t) & c_{22}(t) & ... & c_{2M}(t) \\ c_{31}(t) & c_{32}(t) & ... & c_{3M}(t) \\ ... & ... & ... & ... \\ c_{N1}(t) & c_{N2}(t) & ... & c_{NM}(t) \\ r_1(t) & r_2(t) & ... & r_M(t) \end{pmatrix} \end{array} \tag{4.1}$$

A principal desvantagem do *AdaBoost* é o fato de que as amostras inseridas no classificador não possuem nenhuma relação temporal entre elas, ou seja, a *Amostra* 1 não necessariamente é imediatamente anterior à *Amostra* 2 no tempo. Isto não é conveniente para este sistema de classificação de melhores momentos, pois para análise de algumas características, tais como as referentes ao áudio e movimento de câmera, os valores das amostras vizinhas são extremamente importantes.

Por este motivo, é interessante criar um mecanismo que adapte o *AdaBoost* a esta necessidade. Como o *AdaBoost* permite que sejam inseridas tantas características quanto se queira, a idéia consiste basicamente em adicionar novas linhas de característica só que deslocadas no tempo, como exibido na Equação (4.2).

O classificador *AdaBoost* foi implementado através do *GML AdaBoost Matlab Toolbox* que pode ser encontrado em VEZHNEVETS [97]. Este *Toolbox* possui duas configurações principais, sendo elas a profundidade da árvore interna, que neste trabalho será igual a três por motivos de desempenho, e o número de iterações, que é o número de vezes que os pesos do *AdaBoost* serão ajustados. Tanto o valor ideal do número de iterações quanto o valor ideal do número de amostras passadas e futuras anteriormente citadas serão investigados no Capítulo 5.

Além dos parâmetros destacados, o *Toolbox* fornece três variações do *AdaBoost*: *Real*, *Gentle* e *Modest* que podem ser melhor compreendidos em SCHAPIRE e SINGER [98], FRIEDMAN *et al.* [99] e VEZHNEVETS e VEZHNEVETS [100], respectivamente. O desempenho e comparação entre estas derivações também serão estudados no Capítulo 5. A formulação do classificador *AdaBoost* pode ser vista com mais detalhes no Apêndice A.

Por fim, como o classificador oferece o resultado quadro a quadro, é interessante aplicar um filtro de medianas neste resultado com o objetivo de amenizar respostas espúrias do classificador no meio de outras respostas. Isto se aplica no caso deste trabalho, pois é considerado impraticável que poucos quadros de vídeo que não sejam classificados como "bom momento" estejam entre diversos outros quadros classificados como "bom momento".

$$
\begin{array}{l}
 \\
\textit{Característica } 1 \\
\textit{Característica } 1 \\
\textit{Característica } 1 \\
\textit{Característica } 1 \\
\textit{Característica } 1 \\
\textit{Característica } 1 \\
\textit{Característica } 1 \\
\textit{Característica } 2 \\
\textit{Característica } 2 \\
\textit{Característica } 2 \\
\textit{Característica } 2 \\
\textit{Característica } 2 \\
\textit{Característica } 2 \\
\textit{Característica } 3 \\
\textit{Característica } 3 \\
\textit{Característica } 3 \\
\textit{Característica } 3 \\
\textit{Característica } 3 \\
\textit{Característica } 3 \\
\ldots \\
\textit{Característica } N \\
\textit{Característica } N \\
\textit{Característica } N \\
\textit{Característica } N \\
\textit{Característica } N \\
\textit{Rótulo}
\end{array}
\begin{pmatrix}
\textit{Amostra } 1 & \textit{Amostra } 2 & \ldots & \textit{Amostra } M \\
\ldots & \ldots & \ldots & \ldots \\
c_{11}(t+2) & c_{12}(t+2) & \ldots & c_{1M}(t+2) \\
c_{11}(t+1) & c_{12}(t+1) & \ldots & c_{1M}(t+1) \\
c_{11}(t) & c_{12}(t) & \ldots & c_{1M}(t) \\
c_{11}(t-1) & c_{12}(t-1) & \ldots & c_{1M}(t-1) \\
c_{11}(t-2) & c_{12}(t-2) & \ldots & c_{1M}(t-2) \\
\ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
c_{21}(t+2) & c_{22}(t+2) & \ldots & c_{2M}(t+2) \\
c_{21}(t+1) & c_{22}(t+1) & \ldots & c_{2M}(t+1) \\
c_{21}(t) & c_{22}(t) & \ldots & c_{2M}(t) \\
c_{21}(t-1) & c_{22}(t-1) & \ldots & c_{2M}(t-1) \\
c_{21}(t-2) & c_{22}(t-2) & \ldots & c_{2M}(t-2) \\
\ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
c_{31}(t+2) & c_{32}(t+2) & \ldots & c_{3M}(t+2) \\
c_{31}(t+1) & c_{32}(t+1) & \ldots & c_{3M}(t+1) \\
c_{31}(t) & c_{32}(t) & \ldots & c_{3M}(t) \\
c_{31}(t-1) & c_{32}(t-1) & \ldots & c_{3M}(t-1) \\
c_{31}(t-2) & c_{32}(t-2) & \ldots & c_{3M}(t-2) \\
\ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
c_{N1}(t+2) & c_{N2}(t+2) & \ldots & c_{NM}(t+2) \\
c_{N1}(t+1) & c_{N2}(t+1) & \ldots & c_{NM}(t+1) \\
c_{N1}(t) & c_{N2}(t) & \ldots & c_{NM}(t) \\
c_{N1}(t-1) & c_{N2}(t-1) & \ldots & c_{NM}(t-1) \\
c_{N1}(t-2) & c_{N2}(t-2) & \ldots & c_{NM}(t-2) \\
r_1(t) & r_2(t) & \ldots & r_M(t)
\end{pmatrix}
\tag{4.2}
$$

4.3 Base de Dados

Como já dito na seção anterior, a etapa de treinamento é essencial para a construção de um classificador, o que mostra a importância do processo de montagem da base de dados que será utilizada durante este treinamento. Em teoria, esta base de dados deve conter uma larga gama de tipos de lances, horários e estádios que podem ocorrer em uma partida de futebol. Além disso, para o caso de transmissões televisivas também é importante generalizar a base de dados para os narradores, e as regras de produção da transmissão das diferentes emissoras de TV e campeonatos.

Na prática, contemplar todas essas variações para transmissões televisivas de futebol é uma tarefa muito trabalhosa mas que se feita de maneira adequada pode formar uma base de dados que represente o futebol como ele costuma ser transmitido no Brasil. A Tabela 4.2 mostra todas as 30 transmissões capturadas que procuram representar a variedade necessária em todos aspectos abordados neste trabalho. São eles:

- 5 campeonatos diferentes o que provoca regras de produção de TV diferentes, tais como troca de câmeras e efeitos gráficos;
- 30 seleções e times diferentes o que ocasiona cores de uniformes diferentes;
- 15 estádios diferentes o que resulta em tonalidade do verde do gramado e distâncias das câmeras diferentes;
- partidas durante a tarde, noite e que se iniciam a tarde e terminam a noite o que também causa tonalidades de verde diferentes;
- 4 quatro narradores diferentes causando excitação e características de voz diferentes;
- 86 gols e 421 lances em que há ameaça de gol.

A primeira tarefa a ser feita na montagem da base de dados é a captura das transmissões televisivas que a compõem. Foram capturadas partidas transmitidas pela TV Globo no formato SD contendo 486 linhas e 720 colunas a uma taxa de amostragem de 29.97 quadros por segundo e codificado em MPEG-2 a uma taxa de 6 Mbps. O áudio foi capturado a 48000 Hz e codificado em PCM.

Após a captura, a próxima tarefa na montagem da base de dados foi gerar os rótulos dos lances da partida através da análise visual das transmissões de TV. Além de muito esforço, esta tarefa requer muita atenção, pois é nela em que se define o critério de identificação de cada lance que será adotado. Para alguns casos, esta não é uma tarefa complicada, como no caso de classificação de gols, pois é

Tabela 4.2: Informações dos vídeos que compõem a base de dados. G = Número de gols. LP = Número de lances perigosos.

	Partida	Torneio	Estádio	Narrador	Turno	G	LP
1	Argentina x Alemanha	Copa do Mundo 2010	Green Point	Luís Roberto	Tarde	3	17
2	Argentina x México	Copa do Mundo 2010	Soccer City	Cléber Machado	Noite	4	13
3	Argentina x Nigéria	Copa do Mundo 2010	Ellis Park	Galvão Bueno	Tarde/Noite	1	17
4	Argentina x Coréia do Sul	Copa do Mundo 2010	Soccer City	Cléber Machado	Tarde	5	11
5	Brasil x Chile	Copa do Mundo 2010	Ellis Park	Galvão Bueno	Noite	3	14
6	Brasil x Holanda	Copa do Mundo 2010	Port Elizabeth	Galvão Bueno	Tarde/Noite	3	12
7	Chile x Suíça	Copa do Mundo 2010	Port Elizabeth	Rogério Pinheiro	Tarde/Noite	2	13
8	Dinamarca x Japão	Copa do Mundo 2010	Royal Bafokeng	Luís Roberto	Noite	4	15
9	França x México	Copa do Mundo 2010	Peter Mokaba	Galvão Bueno	Noite	2	12
10	Alemanha x Inglaterra	Copa do Mundo 2010	Free State	Luís Roberto	Tarde/Noite	5	20
11	Alemanha x Espanha	Copa do Mundo 2010	Durban	Galvão Bueno	Noite	1	13
12	Alemanha x Uruguai	Copa do Mundo 2010	Port Elizabeth	Cléber Machado	Noite	5	13
13	Itália x Eslováquia	Copa do Mundo 2010	Ellis Park	Cléber Machado	Tarde/Noite	6	11
14	Holanda x Japão	Copa do Mundo 2010	Durban	Cléber Machado	Tarde	1	11
15	Holanda x Eslováquia	Copa do Mundo 2010	Durban	Luís Roberto	Tarde/Noite	2	4
16	Portugal x Coréia do Norte	Copa do Mundo 2010	Green Point	Luís Roberto	Tarde	7	15
17	Espanha x Holanda	Copa do Mundo 2010	Soccer City	Galvão Bueno	Noite	0	18
18	Espanha x Portugal	Copa do Mundo 2010	Green Point	Cléber Machado	Noite	1	17
19	Espanha x Suíça	Copa do Mundo 2010	Durban	Luís Roberto	Tarde/Noite	1	18
20	Uruguai x Holanda	Copa do Mundo 2010	Green Point	Cléber Machado	Noite	5	13
21	Uruguai x Coréia do Sul	Copa do Mundo 2010	Port Elizabeth	Luís Roberto	Tarde/Noite	3	18
22	Brasil x Itália	Copa das Confederações 2009	Loftus Versfeld	Galvão Bueno	Noite	3	22
23	Espanha x Estados Unidos	Copa das Confederações 2009	Free State	Cléber Machado	Noite	2	16
24	Atlético Mineiro x Vasco	Campeonato Brasileiro 2010	Mineirão	Luís Roberto	Tarde/Noite	3	15
25	Corinthians x Fluminense	Campeonato Brasileiro 2010	Pacaembu	Luís Roberto	Tarde/Noite	1	15
26	Santos x Vasco	Campeonato Brasileiro 2010	Vila Belmiro	Rogério Pinheiro	Tarde/Noite	4	10
27	Barcelona x Internazionale	Liga dos Campeões 2010	Camp Nou	Galvão Bueno	Noite	2	9
28	Bayern Münich x Internazionale	Liga dos Campeões 2010	Santiago Bernabéu	Galvão Bueno	Tarde/Noite	2	14
29	Cruzeiro x São Paulo	Copa Libertadores 2010	Mineirão	Cléber Machado	Noite	3	10
30	São Paulo x Cruzeiro	Copa Libertadores 2010	Morumbi	Cléber Machado	Noite	2	15

fácil discriminar um lance de gol de outros lances. Entretanto, para a proposta deste trabalho, que é classificar lances como "bom momento", esta tarefa torna-se um pouco mais complexa, uma vez que um lance pode ser considerado como bom momento por uma pessoa enquanto outra não o considera. Além disso, também há subjetividade na seleção do início e fim do trecho de bom momento.

Para este trabalho, foram considerados como "bom momento" os lances onde houve ameaça de gol, como, por exemplo, chutes efetuados na direção do gol ou que passaram perto das traves, momentos em que a bola está na área sob domínio de um jogador de ataque em situação clara de gol, entre outras situações semelhantes. Não foram considerados como "bom momento" lances de faltas duras que resultassem em aplicação de cartão pelo juiz, ou dribles que provocassem euforia na audiência e locutores, entre outras situações que não caracterizassem uma tentativa de gol. Além disso, o início do "bom momento" foi definido como o instante em que a situação de gol fica clara, por exemplo, poucos segundos antes do jogador chutar uma bola que o goleiro defendeu. Já o fim do "bom momento" foi marcado no instante em que ocorre a primeira pausa na locução do narrador logo após ao término da ameaça de gol ou do gol. Porém, a marcação ideal de início e fim do lance não é tão importante quanto a indicação dos melhores momentos existentes na partida, e, consequentemente, o ajuste de início e fim é uma tarefa que pode ser facilmente executada posteriormente pelo operador uma vez que ele possui os principais lances da partida retornados pelo sistema.

Apesar de possuir uma dose de subjetividade, não é crítico utilizar um critério próprio de classificação, pois no modelo tradicional de identificação de melhores momentos, esta subjetividade também é encontrada, pois há mais de um operador realizando esta tarefa na emissora de TV. Portanto, por ser uma tarefa que requer um longo tempo para observação de todas as partidas, ela foi dividida em duas partes, sendo a primeira uma seleção a grosso modo dos lances, e a segunda um ajuste fino do início e fim de cada lance. Essa divisão em partes evita que o critério utilizado para seleção de início e fim dos lances seja alterado ao longo do processo de marcação dos lances.

Por fim, a última tarefa na montagem da base de dados, foi empregar todos os algoritmos desenvolvidos ao longo dos Capítulos 2 e 3 em todos os lances marcados anteriormente a fim de obter os seus respectivos valores de características indicados na Tabela 4.1. Para cada lance de bom momento extraído para a base de dados, foi extraído um de lance normal com a mesma duração. Isto garante que a base de dados seja equilibrada contendo 50% de amostras para cada tipo de lance, o que é interessante para o treinamento de classificadores *AdaBoost*.

4.4 Validação Cruzada

Diferentemente das etapas anteriores, a etapa de validação não é essencial para a construção do classificador, porém é de suma importância para avaliar o seu desempenho. Para o sistema proposto neste trabalho, a validação cruzada é interessante pelo fato de todas as amostras da base de dados participarem tanto como amostra contida no conjunto de treinamento quanto como amostra contida no conjunto de teste.

Para qualquer técnica de validação, a base de dados costuma ser dividida entre conjuntos de treino e teste, o que, para o caso deste trabalho, pode ser feito dividindo as amostras em quadros de vídeo ou em partidas. A vantagem ao utilizar o critério de partidas é que fica mais fácil valer-se das informações de cada partida, tais como narrador, estádio entre outros, para realizar a separação adequada da base de dados nestes dois conjuntos.

Há muitas maneiras de utilizar a base de dados para efetuar a validação cruzada, sendo as mais comuns:

- O *leave-one-out*, onde $N - 1$ amostras compõem o conjunto de treinamento enquanto somente uma amostra é utilizada para a validação. Este processo é feito N vezes de forma que todas as amostras sejam amostras de validação; e

- O *k-fold*, ilustrado na Figura 4.1, consiste em dividir a base de dados em k grupos de amostras, onde $k-1$ grupos são utilizados no treinamento enquanto o outro grupo é utilizado na validação. Da mesma forma que no *leave-one-out*, este processo é feito k vezes com o intuito de que todos os *folds* participem como *fold* de validação.

O método de validação *k-fold* se destaca como mais vantajoso pelo simples fato de ser menos custoso computacionalmente, uma vez que a etapa de treinamento conterá menos partidas que no caso do *leave-one-out*. Além de utilizar todas as amostras no treinamento e validação, o fato de realizar este processo k vezes faz com que o resultado final da validação obtenha uma média e desvio padrão das medidas de acerto. O desvio padrão pode indicar a confiabilidade do classificador.

A base de dados foi dividida em oito *folds* de acordo com o visto na Tabela 4.3, onde somente os sete primeiros serão utilizados na validação cruzada. O *fold* 8, pelo fato das partidas terem sido transmitidas por um narrador não presente nos demais *folds*, será utilizado somente para uma validação final do sistema construído com o treinamento realizado com os *folds* 1-7.

Os *folds* que serão utilizados na validação cruzada foram formados de forma a serem heterogêneos, pois, assim, cada *fold* terá mais de um narrador presente, e partidas com times e campeonatos diferentes.

Tabela 4.3: Base de dados dividida em *folds*.

	Partida	Narrador	Turno
Fold 1	Argentina x Alemanha	Luís Roberto	Tarde
	Espanha x Portugal	Cléber Machado	Noite
	Corinthians x Fluminense	Luís Roberto	Tarde/Noite
	Barcelona x Internazionale	Galvão Bueno	Noite
Fold 2	Dinamarca x Japão	Luís Roberto	Noite
	Holanda x Japão	Cléber Machado	Tarde
	Uruguai x Coréia do Sul	Luís Roberto	Tarde/Noite
	Espanha x EUA	Cléber Machado	Noite
Fold 3	Alemanha x Inglaterra	Luís Roberto	Tarde/Noite
	Itália x Eslováquia	Cléber Machado	Tarde/Noite
	Espanha x Holanda	Galvão Bueno	Noite
	Cruzeiro x São Paulo	Cléber Machado	Noite
Fold 4	Brasil x Chile	Galvão Bueno	Noite
	França x México	Galvão Bueno	Noite
	Alemanha x Uruguai	Cléber Machado	Noite
	Holanda x Eslováquia	Luís Roberto	Tarde/Noite
Fold 5	Argentina x Nigéria	Galvão Bueno	Tarde/Noite
	Alemanha x Espanha	Galvão Bueno	Noite
	Portugal x Coréia do Norte	Luís Roberto	Tarde
	Atlético Mineiro x Vasco	Luís Roberto	Tarde/Noite
Fold 6	Argentina x Coréia do Sul	Cléber Machado	Tarde
	Espanha x Suíça	Luís Roberto	Tarde/Noite
	Uruguai x Holanda	Cléber Machado	Noite
	Brasil x Itália	Galvão Bueno	Noite
Fold 7	Argentina x México	Cléber Machado	Noite
	Brasil x Holanda	Galvão Bueno	Tarde/Noite
	Bayern Münich x Internazionale	Galvão Bueno	Tarde/Noite
	São Paulo x Cruzeiro	Cléber Machado	Noite
Fold 8	Chile x Suíça	Rogério Pinheiro	Tarde/Noite
	Santos x Vasco	Rogério Pinheiro	Tarde/Noite

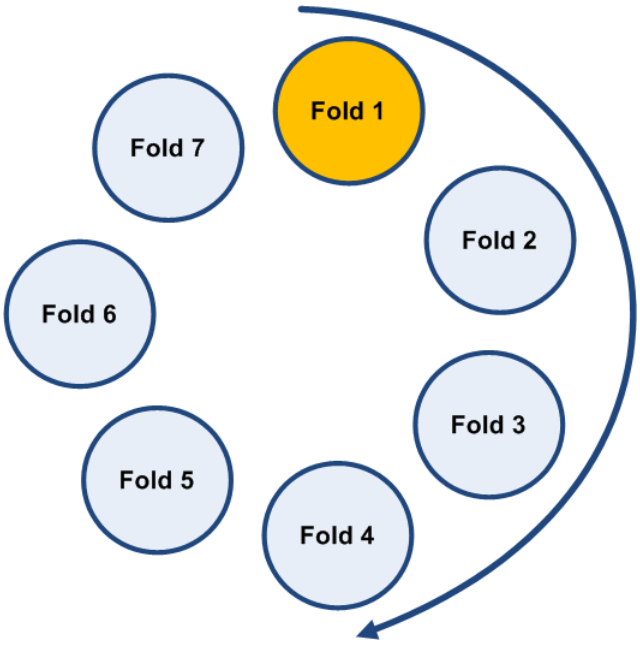


Figura 4.1: Esquema de revezamento do *fold* que é utilizado para validação do sistema.

4.5 Roteiro Experimental

Além da etapa de validação já descrita anteriormente, será acrescida uma nova etapa que tem o objetivo de gerar estatísticas capazes de quantificar a importância das características de áudio e vídeo assim como os processamentos inseridos no uso do classificador *AdaBoost*. Para aplicar esta etapa, a Tabela 4.4 exibe o roteiro de testes que será executado, onde a cada teste, a validação cruzada proposta na Seção 4.4 é aplicada.

Um dos objetivos destes testes é comparar as estatísticas de classificação do sistema contendo somente características de vídeo (Testes 1 a 4), somente características de áudio (Testes 5 a 8) e contendo as de vídeo e áudio simultaneamente (Testes 9 a 12), onde será possível notar a contribuição das características se utilizadas separadamente ou em conjunto. O segundo objetivo, é avaliar a contribuição do uso de amostras passadas e futuras e do filtro de medianas no classificador *AdaBoost* para os três cenários já citados.

4.6 Conclusões

O objetivo deste capítulo foi estudar como as características de áudio e vídeo apresentadas nos Capítulos 2 e 3, respectivamente, podem ser combinadas a fim de

Tabela 4.4: Roteiro de testes indicando fatores considerados em cada um deles.

	Cor	Movimento de Câmera	*Pitch*	Energia	Passado/Futuro	Filtro Medianas
Teste 1	X	X				
Teste 2	X	X				X
Teste 3	X	X			X	
Teste 4	X	X			X	X
Teste 5			X	X		
Teste 6			X	X		X
Teste 7			X	X	X	
Teste 8			X	X	X	X
Teste 9	X	X	X	X		
Teste 10	X	X	X	X		X
Teste 11	X	X	X	X	X	
Teste 12	X	X	X	X	X	X

obter o melhor resultado de classificação para sistemas de melhores momentos de transmissões televisivas de partidas de futebol.

A Seção 4.1 foi responsável por descrever brevemente o que tem sido feito na literatura a respeito de combinação de características para o tipo de sistema proposto neste trabalho.

Após isto, a Seção 4.2 descreveu como deve ser a construção de um classificador e apresentou o *AdaBoost* como solução para sistemas de melhores momentos pelo fato de ser possível construir um classificador forte a partir de diversas características consideradas fracas. Além disso, foram propostas operações de pré- e pós-processamento com o objetivo de adaptar o *AdaBoost* ao cenário encontrado neste trabalho, o que por definição não é contemplado pelo *AdaBoost*.

Em seguida, a Seção 4.3 demonstra a construção da base de dados que será utilizada para as etapas de treinamento e validação do sistema.

Posteriormente, a Seção 4.4 apresentou um método de validação cruzada, onde todas as amostras contidas na base de dados são utilizadas tanto nas etapas de treinamento quanto na etapa de validação. Além disso, o resultado da validação cruzada é interessante, pois além da medida de acerto do classificador, obtém-se também uma medida de confiabilidade do classificador.

Por fim, a Seção 4.5 apresenta uma nova etapa inserida na construção do clasificador que tem como intuito avaliar a importância das características de áudio e vídeo, além do pré- e pós-processamento adaptados ao *AdaBoost*.

Assim, este capítulo desenvolveu uma metodologia para que as características de áudio e vídeo sejam combinadas e para que o próximo capítulo seja capaz de obter

e analisar os resultados interessantes para a validação do sistema construído neste trabalho.

Capítulo 5

Resultados Experimentais

O capítulo 4 demonstrou como foi feita a construção do classificador *AdaBoost* que processa todas as características extraídas da transmissão televisiva de uma partida de futebol. Este capítulo tem como principal objetivo aplicar a metodologia descrita anteriormente, avaliar seus resultados, definir como deve ser o classificador, para, por fim, medir o desempenho do sistema na visão operacional.

A Seção 5.1 descreve medidas de avaliação que foram utilizadas com o intuito de quantificar os resultados do classificador. Em seguida, a Seção 5.2 define os parâmetros importantes para a construção do classificador. A Seção 5.3 faz uso dos parâmetros definidos para aplicar o roteiro experimental apresentado no capítulo anterior, para, então, destacar a importância das características de áudio e vídeo, assim como o pré- e pós-processamentos inseridos no classificador. Na sequência, a Seção 5.4 valida o sistema utilizando uma visão do usuário. Por fim, a Seção 5.5 compara os resultados obtidos ao longo do capítulo com outros algoritmos similares encontrados na literatura.

5.1 Medidas de Avaliação

De forma a analisar os resultados gerados pela validação de maneira mais criteriosa, foram utilizadas três medidas para avaliar o desempenho do sistema. A Figura 5.1 exibe uma representação gráfica do que essas três medidas significam no cenário de transmissões de partidas de futebol.

A primeira delas, apresentada na Equação (5.1), denominada como taxa de precisão TP, mede a porção dos bons momentos listados pelo sistema que efetivamente são bons momentos. A segunda, representada pela Equação (5.2) denominada como taxa de *recall* TR, significa o percentual dos bons momentos existentes no trecho de transmissão validado que realmente foram encontrados pelo sistema. A última, indicada na Equação (5.3), denominada como taxa de resumo TS, significa o percentual da duração da partida que foi marcada como pertencente a bons momentos.

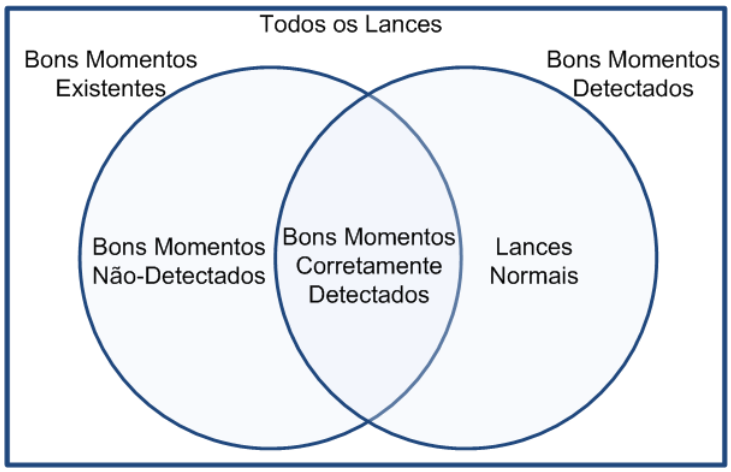


Figura 5.1: Diagrama ilustrando as medidas úteis para avaliar o desempenho do classificador. Em um sistema de sumarização ideal, onde todos os bons momentos, e somente eles, são detectados, os círculos seriam sobrepostos.

$$TP(\%) = \frac{\text{Bons Momentos Corretamente Detectados}}{\text{Bons Momentos Detectados}} , \tag{5.1}$$

$$TR(\%) = \frac{\text{Bons Momentos Corretamente Detectados}}{\text{Bons Momentos Existentes}} , \tag{5.2}$$

$$TS(\%) = \frac{\text{Duração Bons Momentos Detectados}}{\text{Duração Todos Lances}} . \tag{5.3}$$

Apesar da taxa de precisão por vezes ser a medida mais importante, por evidenciar o acerto do sistema, para sistemas de detecção de melhores momentos de uma transmissão de futebol é mais interessante que todos os bons momentos existentes na transmissão sejam encontrados pelo sistema. Em outras palavras, para que o usuário confie no sistema é importante que este atinja 100% de taxa de *recall*, pois, assim, todos os bons momentos da transmissão estarão no resumo gerado, mesmo que a taxa de precisão não esteja próxima aos 100%.

Quando os resultados forem gerados a partir do método de validação cruzada apresentado na Seção 4.4, as taxas de precisão e *recall* terão a média e o desvio padrão dos resultados dos *k-folds* montados a partir da base de dados. No estágio de definição dos parâmetros e avaliação inicial do sistema, a taxa de resumo TS não será utilizada por não ser tão importante nesta questão, pois ela é útil somente para demonstrar em quanto o vídeo original completo foi resumido.

5.2 Definição de Parâmetros

Na Seção 4.2 foi visto que para a construção do classificador baseado no *AdaBoost* é necessária a definição de alguns parâmetros, tais como o tipo de classificador *AdaBoost*, o número It de iterações, o número PF de amostras passadas/futuras e o tamanho M do filtro de medianas.

Por não haver referência para o uso do *AdaBoost* da maneira proposta por este trabalho, os valores ótimos dos parâmetros serão encontrados de forma experimental. Para isto, foi criada uma rotina que provocasse a validação cruzada de vários cenários, cada um contendo diferentes valores dos diversos parâmetros, para que ao fim fosse obtida a configuração de melhores as taxas de precisão e *recall*.

A execução desta rotina gerou muitos resultados, dentre eles os exibidos na Tabela 5.1, onde são vistas as taxas de precisão e *recall* ótimas encontradas, e seus parâmetros associados. Como já dito anteriormente, para sistemas de melhores momentos de transmissões de futebol, é mais interessante que todos os bons momentos existentes sejam listados pelo classificador do que todos os listados pelo classificador estejam corretos. Por isso, optou-se por utilizar os parâmetros definidos pela melhor taxa de *recall*.

Tabela 5.1: Medidas de desempenho e parâmetros ótimos encontrados após a execução da rotina.

	TP Ótima	TR Ótima
Média	92.61%	91.92%
Desvio Padrão	1.78%	2.87%
Tipo	*Gentle*	*Gentle*
It	31	33
PF	30	30
M	15	9
Recall Médio	91.82%	-
Precisão Média	-	92.59%

Os cenários apontados pela TP ótima e pela TR ótima estão muito próximos, mas, apesar da taxa de precisão, 92.59%, associada aos valores de parâmetros oriundos da taxa de *recall* ótima ser menor do que a taxa de precisão ótima, 92.61%, os valores são muito próximos, estando dentro dos limites definidos pelo desvio padrão da taxa de precisão ótima, 1.78%. Portanto, supõe-se que utilizar os parâmetros definidos pela taxa de *recall* ótima não fará com que a taxa de precisão diminua substancialmente.

Vale ressaltar que o fato da taxa de *recall* não chegar aos 100% não é considerado crítico neste momento, pois a validação feita nesta etapa do desenvolvimento do

sistema, considera como uma não-marcação adequada o caso em que classificador começa a indicar temporalmente um bom momento anteriormente ou posteriormente ao rótulo determinado durante a construção da base de dados exibido na Seção 4.3. A Seção 5.4 trará as taxas práticas do sistema, em outras palavras, as taxas de precisão e *recall* na visão do usuário do sistema.

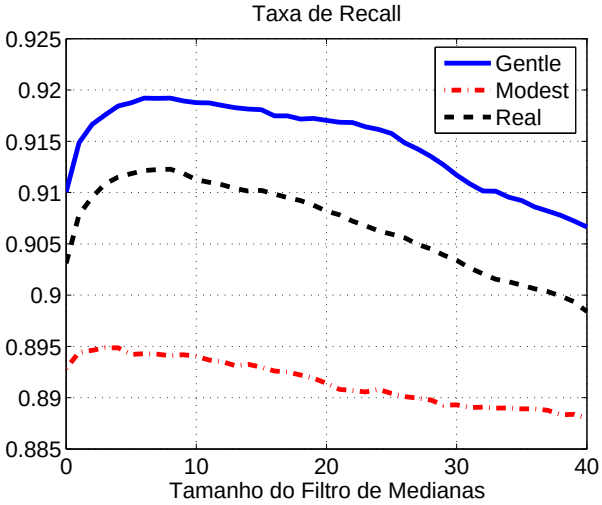
Como forma de ilustração, a Figura 5.2 exibe gráficos que mostram a variação da taxa de *recall* para cada parâmetro quando os demais estão fixados em seus valores ótimos. Em todos os casos, é fácil perceber que *Gentle* e *Real* são os tipos de classificador *AdaBoost* que atingem as melhores taxas, com o primeiro se destacando ligeiramente, principalmente para a definição do tamanho do filtro de medianas.

O gráfico da Figura 5.2a mostra que o tamanho do filtro de mediana ótimo M está entre 5 e 10. Como a aplicação do filtro de medianas não é um processo custoso computacionalmente, não há necessidade de escolher o menor tamanho possível, o que torna o valor ótimo $M = 9$ interessante. O gráfico também mostra que a inserção do filtro de medianas provoca o aumento de aproximadamente 1% sobre os 91% na taxa de *recall* atingidos anteriormente.

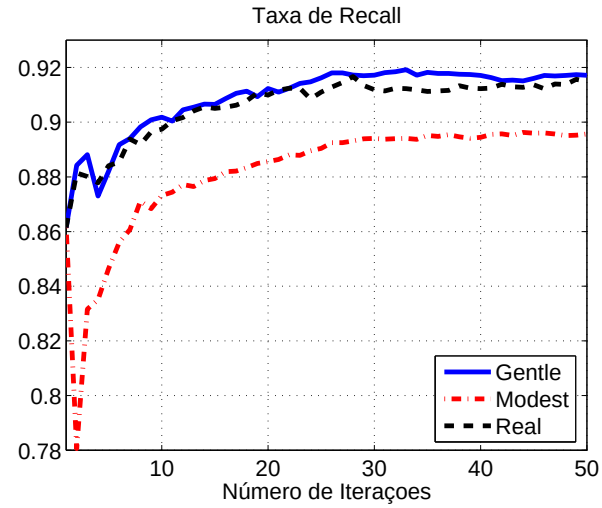
Apesar do valor do número de iterações ótimo It ter sido definido como 33, o gráfico da Figura 5.2b indica que a taxa de *recall* converge após $It = 30$. Isto nos faz crer que qualquer valor acima de $It = 30$ trará resposta similar ao sistema, mas pelo fato do número de iterações influenciar diretamente no custo computacional do sistema é interessante defini-lo com o menor valor possível no caso, $It = 30$. Este valor leva a uma taxa média de *recall* de 91.72% que é próxima a taxa ótima de 91.92%. É também possível verificar no gráfico que após as 30 iterações no treinamento do *AdaBoost*, a taxa de *recall* sofre um aumento de cerca de 6%.

Por último, o gráfico da Figura 5.2c mostra a evolução da taxa de *recall* de acordo com o número de amostras passadas/futuras. Neste experimento, é fácil ver que o melhor valor para este parâmetro está ao redor de $PF = 30$ amostras passadas/futuras. Como a cada inserção de amostra passada/futura, o conjunto de treinamento é acrescido de duas características para cada característica já existente, o cálculo do número ótimo PF foi variado de um em um até o número 15 e de cinco em cinco após isso para reduzir o tempo de treinamento. No gráfico, nota-se também que a introdução de características que representam amostras passadas e futuras no *AdaBoost* ocasionou um aumento de aproximadamente 10% na taxa de *recall* do classificador, o que pode ser considerado excelente dado que a taxa inicial já era maior do que 80%.

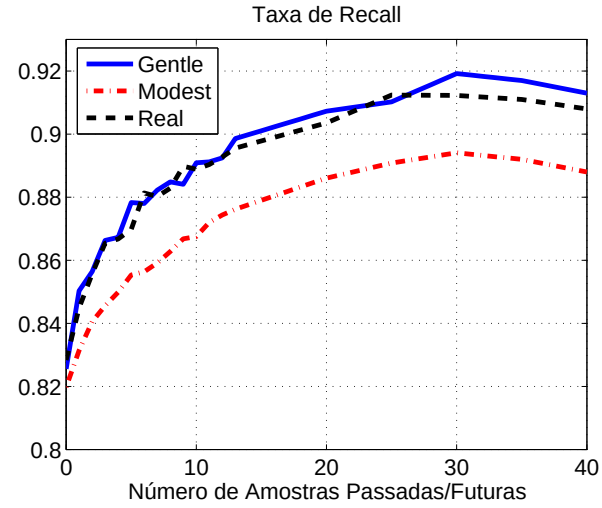
Os gráficos nos levaram a concluir que a escolha correta do número de amostras passadas e futuras traz maiores ganhos do que as iterações do classificador e do filtro de medianas, o que comprova a necessidade de incorporar a informação temporal que o *AdaBoost* inicialmente não prevê.



(a)



(b)



(c)

Figura 5.2: Ilustrações gráficas dos valores ótimos dos parâmetros tamanho do filtro de medianas (a), número de iterações (b) e o número de amostras passadas/futuras (c).

5.3 Roteiro Experimental

Na seção anterior, já foi possível notar a importância de algumas etapas do dimensionamento do sistema. Esta seção também destacará a importância das já citadas e de outras etapas mas aplicando o roteiro experimental descrito na Seção 4.5. A influência de cada etapa para o sistema final é visto na Tabela 5.2.

Tabela 5.2: Taxas de precisão e *recall* resultantes do roteiro experimental. A = Áudio. V = Vídeo. FM = Filtro de Medianas. PF = Amostras Passadas/Futuras.

Teste	Resumo	Taxa de Precisão		Taxa de *Recall*	
		Média	Desvio Padrão	Média	Desvio Padrão
1	V	61.21%	2.30%	63.69%	11.29%
2	V + FM	61.91%	2.45%	64.75%	11.99%
3	V + PF	65.70%	2.11%	66.40%	9.24%
4	V + FM + PF	66.34%	2.41%	67.23%	9.73%
5	A	83.58%	1.86%	80.24%	4.94%
6	A + FM	86.56%	2.32%	82.27%	5.81%
7	A + PF	90.71%	2.30%	90.04%	4.02%
8	A + FM + PF	91.78%	2.65%	90.61%	4.40%
9	A + V	83.37%	1.41%	80.88%	3.73%
10	A + V + FM	85.95%	1.96%	82.50%	4.66%
11	A + V + PF	91.22%	1.60%	90.92%	2.90%
12	A + V + FM + PF	92.55%	1.68%	91.74%	3.19%

A Tabela 5.2 mostra os resultados do roteiro experimental, mas para uma visão mais geral da importância de cada estágio do sistema, a Tabela 5.3 é mais adequada. Nela, os valores indicados são as médias dos resultados referentes a cada etapa. Por exemplo, os valores do sistema que somente contém características de vídeo são as médias dos experimentos 1-4 da Tabela 5.2.

A primeira avaliação que pode ser feita, é em relação às características que compõem o sistema. Ao utilizar somente características de vídeo, percebe-se que as médias das taxas de precisão e de *recall* estão bem abaixo do que quando se utiliza somente as características de áudio. Isto era esperado, pois as características de áudio descrevem a ocorrência da aplicação de emoção na voz, o que aponta diretamente que algo de interessante aconteceu. Enquanto isso, as características de vídeo descrevem somente ocorrências que não estão diretamente associadas a bons momentos, como a cor dominante da imagem e a troca de tipo de câmera em uso.

Pelo mesmo motivo, ao utilizar tanto as características de áudio quanto as de vídeo, observa-se que as taxas TP e TR são pouco maiores do que quando somente as características de áudio são utilizadas. Este fato indica que o objetivo inicial das características de vídeo, que é auxiliar o classificador nos momentos em que somente

Tabela 5.3: Resultados do roteiro experimental separado em etapas.

Características			
	Experimentos	Taxa de Precisão	Taxa de *Recall*
V	1-4	63.79%	65.51%
A	5-8	88.15%	85.65%
A + V	9-12	88.27%	86.51%
Filtro de Medianas			
	Experimentos	Taxa de Precisão	Taxa de *Recall*
Sem	1,3,5,7,9,11	79.31%	78.70%
Com	2,4,6,8,10,12	80.85%	79.85%
Amostras Passadas/Futuras			
	Experimentos	Taxa de Precisão	Taxa de *Recall*
Sem	1,2,5,6,9,10	77.10%	75.73%
Com	3,4,7,8,11,12	83.05%	82.83%

o áudio não é suficiente, está sendo alcançado.

Em seguida, foi avaliada a importância da adoção do filtro de medianas no sistema. O objetivo do filtro de medianas era somente reduzir a ocorrência de casos onde somente um quadro é apontado como um rótulo enquanto os seus vizinhos são apontados com o rótulo contrário, o que considera-se impossível devido ao fato do quadro possuir somente 33.3 milissegundos de duração. Por isso, o modesto aumento nas médias das taxas TP e TR é considerado normal pelo fato de tratar casos especiais e de curta duração do sistema.

Por fim, foi avaliado o ganho que as amostras passadas e futuras inseridas no classificador causariam no sistema. Como normalmente dados extraídos de sinais de áudio e vídeo contém forte relação temporal, já era esperado, e havia sido indicado na seção anterior, que este estágio causasse um aumento significativo nas taxas TP e TR.

Pelos resultados mostrados, pode-se concluir, então, que o uso das características de áudio aliada às características de vídeo, somada às etapas de pré- e pós-processamento constituem o melhor sistema de classificação de melhores momentos que poderia se obter com os recursos apresentados neste trabalho.

5.4 Validação Operacional

Até o momento, pelo fato de serem necessários uma quantidade massiva de testes, foi utilizada a validação automática da base de dados por não ser necessária a intervenção humana para verificar se a saída do classificador está correta. Porém, pelo fato deste tipo de sistema ter uma dose de subjetividade tanto para considerar

lances como bom momento, quanto para delimitar início e fim dos lances, validar o resultado automaticamente não é adequado para se obter resultados de operação do sistema. Isto porque para o usuário alguns erros da validação automática são aceitáveis.

A Figura 5.3 ilustra como a validação operacional trata alguns casos de erro da validação automática.

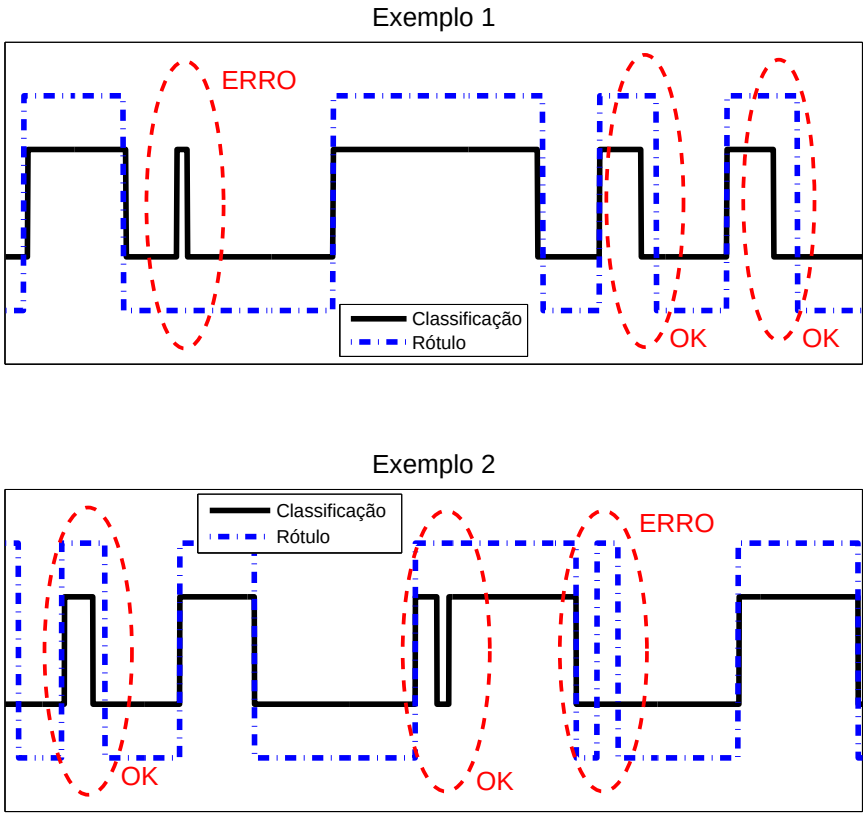


Figura 5.3: Exemplos de casos de erros relevados e não relevados na validação operacional.

Por exemplo, no primeiro caso do Exemplo 1 contido na Figura 5.3, alguns quadros isolados foram classificados como bom momento em uma região onde o rótulo indica lances normais; esta classificação também é considerada um erro para a validação operacional, pois naquela região realmente não há indício de bom momento. Porém, este não é um erro de *recall* pois ele não deixou de classificar um bom momento ocorrido. Nos exemplos seguintes a classificação parou de indicar bom momento antes do rótulo fazê-lo, o que para a validação operacional não é um caso de erro, pois o bom momento ocorrido naquele trecho já foi sinalizado anteriormente, e, desta forma, o usuário o verá. O primeiro caso do Exemplo 2 exibido na Figura 5.3 também não é considerado erro, pelo mesmo motivo dos casos anteriores.

Já no segundo caso do Exemplo 2 ocorre uma classificação de lance normal onde o rótulo indica que se trata de um bom momento. Porém, como a sua volta os demais quadros foram classificados corretamente, estes quadros mal classificados são tolerados pelo operador, mesmo sabendo que ele transforma um único bom momento em dois. Entretanto, o último caso mostra um erro até para a validação operacional, pois nenhum quadro indicou bom momento. Neste caso, além de erro de precisão do sistema, há também o erro de *recall* pois o usuário não saberá da existência deste lance de interesse já que não irá analisar nenhum dos seus quadros de vídeo.

De posse da validação operacional comentada e ilustrada na Figura 5.3, é possível realizar a validação cruzada nos *folds* definidos na Tabela 4.3 aplicando tanto a abordagem automática quanto a operacional.

Tabela 5.4: Resultados das validações automática e operacional feitas nos *folds* da base de dados.

	Automática		Operacional	
Fold	TP (%)	TR (%)	TP (%)	TR (%)
1	92.32	91.98	83.52	100
2	95.39	94.83	89.28	100
3	90.24	86.07	86.66	95.78
4	92.30	93.07	79.72	100
5	92.37	89.91	86.74	97.29
6	93.89	90.88	84.37	94.18
7	91.26	95.43	67.24	100
Média	92.55	91.74	82.50	98.17
Desvio Padrão	1.68	3.19	7.37	2.44

Os resultados das validações vistos na Tabela 5.4 mostram que as taxas de precisão TP são maiores na validação automática do que na operacional. Mas, por outro lado, as taxas de *recall* TR chegam bem mais próximo dos 100% no caso da validação operacional, e isto é muito mais desejável para um sistema que não quer perder lances de interesse do que uma taxa de precisão alta.

Apesar da validação operacional atingir cerca de 98% de taxa de *recall*, este resultado não poderia ser considerado idealmente bom, pois, ainda assim, alguns bons momentos estariam sendo perdidos pelo classificador. Porém, ao analisar os trechos que o classificador perdeu, reparou-se que nenhum deles foi um lance que decididamente seria indicado como lance de perigo por todos usuários, ou seja, foi considerado bom momento durante a definição dos rótulos na base de dados, mas se não fosse não haveria problema devido a subjetividade atribuída a esta tarefa. Além disso, houve lances em que o locutor não deu muita atenção para o que ocorria em

campo, e, por consequência, não aplicou emoção na voz como em outras situações. Entretanto, apesar de alguns lances perdidos, nenhum deles tratava-se de um gol ou de um lance de extremo perigo.

Estas validações foram aplicadas somente em trechos selecionados de cada transmissão contida na base de dados. Com o intuito de avaliar o desempenho do sistema no mundo real, o último experimento consistiu em aplicar ao sistema todos os vídeos completos, inclusive os vídeos 7 e 28 que não participaram do treinamento. A Tabela 5.5 mostra os todos os resultados desse experimento.

Apesar de gols não terem sido perdidos, taxas de *recall* altas, pode-se pensar que os resultados não são satisfatórios devido à baixa taxa de precisão alcançada. Entretanto, ao observar as taxas TS vê-se que mesmo com uma taxa de precisão em torno de 17% em média, o usuário só precisará checar cerca de 12% do vídeo total, o que demonstra uma importante diminuição do montante de vídeo a ser analisado.

Ao analisar os diversos lances normais apontados como bom momento pelo classificador, notou-se que em muitos deles há uma variação na emoção da voz, mesmo que discreta, e como nesses casos a câmera panorâmica com visão do campo de jogo era utilizada, o sistema se confundiu. Isto ocorreu devido ao fato das características de vídeo não serem tão conclusivas quanto as de áudio, fazendo com que não ajudassem o classificador a identificar estes momentos como lance normal.

Entretanto, como o sistema se apresentou confiável, pois em média 97% dos bons momentos estão inclusos no resumo, onde os que não foram inclusos certamente não são os lances mais importantes da partida, o trabalho do usuário se resumirá a observar somente uma parte do vídeo, em média quase 11 minutos, ao invés dos 90 tradicionais e rapidamente poderá descartar todos os trechos de lance normal que o sistema indicou. Assim, os resultados deste trabalho podem ser considerados bons, pois o objetivo de resumir significativamente a transmissão contendo os lances mais importantes da partida foi alcançado.

Outro ponto interessante, foi o fato de que os vídeos 7 e 28, mesmo não tendo participado do treinamento e contendo um locutor diferente dos demais, tiveram resultados próximos ao da média de todos os outros que participaram do treinamento. Isto mostra que o sistema atingiu também o objetivo de ser genérico, principalmente na questão do locutor, pois temia-se que ele ficasse dependente aos narradores que participaram do treinamento do sistema.

5.5 Comparação com Outros Trabalhos

Apesar dos algoritmos propostos na literatura variarem bastante de caso para caso, tornando difícil uma comparação justa, a seguir faremos uma comparação do método proposto com outros algoritmos. Entretanto, deve ser levado em conta que

Tabela 5.5: Resultados da validação operacional feita nos vídeos completos da base de dados. Em negrito os vídeos que não participaram do treinamento do sistema. LP = Número de lances de perigo perdidos na classificação. G = Número de gols perdidos na classificação.

Partida	LP	G	TP (%)	TR (%)	TS (tempo)	TS (%)
1	1	0	20.21	95.00	10:39	13.62
2	0	0	13.38	100	15:37	16.42
3	1	0	13.38	94.40	10:24	10.94
4	1	0	14.58	93.33	11:34	12.42
5	0	0	6.39	100	14:42	15.77
6	1	0	11.29	93.33	12:37	13.37
7	**0**	**0**	**12.29**	**100**	**10:15**	**10.77**
8	1	0	18.55	94.73	9:16	9.64
9	0	0	14.73	100	5:39	6.01
10	0	0	25.00	100	12:23	13.30
11	2	0	28.20	84.60	2:16	2.40
12	0	0	19.14	100	11:06	11.77
13	0	0	17.89	100	11:39	11.92
14	0	0	15.00	100	6:24	6.80
15	0	0	26.08	100	1:45	5.81
16	1	0	22.58	95.45	6:58	12.69
17	3	0	35.71	83.33	2:34	2.70
18	1	0	13.17	94.44	12:48	13.57
19	0	0	21.68	100	7:35	8.10
20	2	0	19.73	88.23	12:27	12.85
21	0	0	22.58	100	10:09	10.79
22	0	0	17.98	100	10:54	11.57
23	0	0	10.65	100	18:53	20.59
24	0	0	17.82	100	7:46	8.30
25	1	0	15.38	93.33	7:37	8.08
26	0	0	12.71	100	11:15	12.10
27	0	0	5.85	100	23:35	28.32
28	**0**	**0**	**12.31**	**100**	**17:14**	**18.25**
29	0	0	8.96	100	12:27	13.05
30	0	0	12.50	100	20:22	21.93
Média	0.46	0	16.85	97.00	10:57	12.12

os resultados destas comparações, apesar de servirem como uma referência, devem ser interpretados com cautela. Em EKIN *et al.* [101], os autores propõem diversos métodos de detecção de características de baixo nível do vídeo para resumir de três maneiras o vídeo de futebol, sendo elas: todos os segmentos em câmera lenta, todos os gols, e todos os segmentos de câmera lenta relacionados a algum objeto específico. EKIN *et al.* [101] foram capazes de resumir o tempo total da partida em 4.68% com uma taxa de precisão de 45%, mas por outro lado, diferentemente deste trabalho, não foram capazes encontrar todos os gols existentes na base de dados, o que é extremamente não desejável para este tipo de algoritmo.

De maneira similar a este trabalho, COLDEFY e BOUTHEMY [17] combinam características de áudio e vídeo para encontrar os melhores momentos da transmissão esportiva. Eles conseguem resumir a partida em 4.5% do tempo total com uma excelente taxa de precisão de 97.50%, mas da mesma forma que EKIN *et al.* [101], gols são perdidos durante a classificação, atingindo uma taxa de *recall* de 88.23% para os gols. A vantagem do algoritmo apresentado por COLDEFY e BOUTHEMY [17] é que ele é um sistema não-supervisionado, ou seja, não necessita de nenhuma etapa de aprendizado.

Já YANG *et al.* [57] apresentam um algoritmo que combina detecção das traves, transição de corte de cena e tipo de cena para encontrar os melhores momentos da partida. Ao retornar somente gols, o sistema atinge 100% de *recall*, e, ao retornar melhores momentos, atinge 88.93% de *recall*.

Por último, ELDIB *et al.* [27] utilizam um inovador detector de *replay* aliado a detectores de traves e grafismos exibindo o placar do jogo para criar um algoritmo capaz de identificar além dos melhores momentos, faltas, cartões entre outros. Eles atingiram 100% de taxa de precisão e *recall* para os gols, mas ao incluir os melhores momentos a taxa de precisão caiu para 86.7% enquanto a de *recall* caiu para 91.7%.

Apesar dos trabalhos citados atingirem taxas de precisão muito maiores e taxas de resumo menores que a deste trabalho, nenhum deles atingiu taxas de *recall* para os melhores momentos da transmissão tão altas quanto as desta dissertação (97.00%). Possivelmente, as taxas de precisão podem melhoradas se mais técnicas de retirada de informação do vídeo, incluindo informação de mais alto nível, comuns nos demais trabalhos, forem aplicadas.

5.6 Conclusões

Este capítulo teve como objetivo aplicar a metodologia proposta no Capítulo 4 para definir o sistema e seus avaliar seus desempenho.

Em primeiro lugar, a Seção 5.1 definiu as medidas que foram usadas ao longo do capítulo para avaliar o desempenho do sistema.

Em seguida, a Seção 5.2 realizou uma validação cruzada de diversos valores dos parâmetros envolvidos no classificador, a fim de obter os seus valores ótimos. Como referência para o valor ótimo, foi visto que a taxa de *recall* é mais interessante do que a taxa de precisão uma vez que o sistema tem como objetivo que a sua saída inclua todos os bons momentos, mesmo que a taxa de precisão seja menor.

Após isto, a Seção 5.3 realizou testes que tinham o intuito de entender qual seria a melhor configuração do sistema de melhores momentos. Para isso, foram feitos testes onde as características de áudio e vídeo, o filtro de medianas, e a informação temporal no classificador eram validados separadamente. Estes testes tornaram possível verificar que estava correta a proposta deste trabalho de combinar características de áudio e vídeo, e adicionar as etapas de pré- e pós-processamento.

Por fim, na Seção 5.4 foi proposta a validação operacional que observa os resultados do sistema de maneira mais adequada para o usuário. Os resultados desta validação indicaram que para o usuário, o sistema atingia desejáveis 98% de taxa de *recall* mesmo com a taxa de precisão mais baixa em torno de 82% para os *folds* do conjunto de treinamento. Apesar de não chegar aos 100%, todos os lances mais importantes da partida, como gols e lance de extremo perigo, estavam contidos no resumo.

Posteriormente, foram aplicados ao sistema todos os vídeos completos da base de dados, e foi visto que a taxa de precisão baixou consideravelmente para a média de cerca de 17%. Entretanto, também foi visto que o sistema se mostrou confiável ao conter em seu resumo todos os lances mais importantes da partida, e que este resumo obteve cerca de 12% do tempo total da partida, tornando a tarefa do usuário muito mais simples do que assistir todos os noventa minutos de partida. Além disso, o sistema se mostrou robusto a vídeos que não fizeram parte do treinamento, inclusive validando de maneira similar vídeos de locutores diferentes.

Capítulo 6

Conclusões

Esta dissertação apresentou um sistema capaz de resumir automaticamente uma partida de futebol transmitida por um emissora de TV em seus melhores momentos para cerca de 12% do tempo total a uma taxa de *recall* de 97%. A proposta do sistema foi encontrar somente os gols e tentativas de gols sem levar em consideração lances como cartões vermelhos, dribles que não resultem em conclusão, entre outros eventos não relacionados ao gol. Com esta finalidade, foram propostos algoritmos para extração de características audiovisuais úteis para a construção de um classificador genérico para transmissões televisivas de futebol.

O Capítulo 2 introduziu algoritmos de extração de características baseados na voz do locutor presente no áudio contido na transmissão de TV, já que é notório que durante momentos de interesse da partida o locutor aplica emoção em sua voz. Para isso, foi proposto um método de estimação da frequência fundamental da voz baseado em cálculos de auto-correlação, e um método baseado na energia de tempo-curto aliado a aplicação de um filtro *Comb* no domínio da frequência para gerar as características de áudio do sistema. Como o áudio da transmissão é uma mistura da voz do locutor com a de comentaristas, repórteres, manifestações da audiência e de todos os participantes da partida, ambos os métodos foram construídos de modo reduzir a interferência de fontes sonoras diferentes da voz. Além disso, para evitar que as variações naturais entre locutores e equipamentos de captura influenciassem nos valores resultantes dos algoritmos, foi proposta uma abordagem que identifica crescimentos locais nos resultados dos métodos já comentados.

Da mesma forma, o Capítulo 3 demonstrou algoritmos de extração de características baseados no vídeo produzido por emissoras de TV. O primeiro algoritmo valeu-se do fato dos campos de futebol serem predominantemente verdes para gerar características que pudessem indicar quando o campo de jogo era ou não o foco da transmissão. O segundo algoritmo, através da análise do movimento de câmera, gerou características que, além da própria informação para onde a câmera está se movendo, contêm informação de que tipo de imagem está sendo exibida em

determinado instante da transmissão, por exemplo, panorâmica. Apesar dessas características não estarem diretamente atreladas aos melhores momentos da partida, elas são úteis para contribuir com o classificador em situações que somente o áudio não é capaz de determinar se dado instante é parte ou não de um bom momento.

Em seguida, o Capítulo 4 apresentou o classificador *AdaBoost* como solução para combinar as características de áudio e vídeo, e, assim, obter os melhores momentos da transmissão. A vantagem decisiva para utilização do *AdaBoost* foi o fato de este ser adequado para casos em que as características que fazem parte do classificador não serem consideradas fortes, como as características de vídeo para o sistema proposto. Por outro lado, o classificador *AdaBoost* não se apresentava suficientemente eficaz para tratar sinais temporais por tratar cada amostra isoladamente. Isto foi facilmente resolvido com uma etapa de pré-processamento que tem o objetivo de adicionar características das amostras passadas e futuras a cada instante de tempo. Além disso, viu-se que seria interessante inserir uma etapa de pós-processamento onde um filtro de medianas é aplicado no resultado da classificação efetuada pelo *AdaBoost* com o intuito de diminuir a incidência de rótulos espúrios em meio a grupos rotulados de outra maneira. Após a definição do classificador, foi apresentado o método de validação cruzada, que ao utilizar todos os vídeos da base de dados tanto para o treinamento quanto para a validação, trouxe resultados mais confiáveis. O Capítulo 4 ainda demonstrou como foi feita a construção da base de dados utilizada para definição de parâmetros, treinamento e validações do sistema.

Finalmente, o Capítulo 5 aplicou a metodologia do capítulo anterior para definir os parâmetros, destacar a importância de cada etapa do sistema, e validá-lo. Para a definição de parâmetros, foram feitas validações cruzadas utilizando-se diversos valores dos parâmetros que compõem o sistema a fim de obter os seus valores ótimos. Após isto, o sistema foi treinado e validado em diversas situações diferentes, onde foi possível verificar que a utilização de todas as características de áudio e vídeo mais as etapas de pré- e pós-processamento resultam no melhor sistema de sumarização de melhores momentos de transmissões televisivas de futebol, como esperado. Também foi visto, que as características de áudio contribuem muito mais para a eficácia do sistema do que as características de vídeo, como era previsto.

Posteriormente, em contraste à validação automática aplicada anteriormente, foi realizada uma validação operacional com o objetivo de obter as taxas de precisão e de *recall* práticas do sistema, ou seja, taxas onde os erros provocados por imprecisões provenientes de marcações de início e fim dos lances de interesse e marcações espúrias fossem descartados. Viu-se que a taxa de *recall* aumentou enquanto a de precisão diminuiu, o que era esperado para um sistema que não deseja perder nenhum lance de interesse. Além disso, foi feita uma validação operacional com todos os vídeos completos da base de dados. Este teste final comprovou que o objetivo

de generalizar o sistema para qualquer transmissão, como as que têm locutores que não participaram do treinamento, foi bem sucedido. Este teste também mostrou que ao utilizar o sistema em vídeos completos, muitos trechos de lance normal eram marcados, reduzindo substancialmente a taxa de precisão para cerca de 17% em média, mas as taxas de *recall* permaneceram altas, em torno de 97% em média, e as taxas de resumo ficaram baixas, próximas a 12% em médida, como desejado.

O sistema se mostrou muito dependente do áudio, pois em alguns casos onde o locutor não aplica emoção, os lances foram perdidos. Em outros, onde o locutor aplica um pouco de emoção e as características de vídeo são favoráveis, lances normais foram apontados como bom momento, provocando uma redução significativa da taxa de precisão.

Alguns outros trabalhos, que possuem objetivos muito próximos ao deste, geram resumos menores com taxas de precisão muito maiores, mas, por outro lado, têm taxas de *recall* menores, por vezes, perdendo lances importantes como gol.

Além disso, os resultados em geral, mostram um sistema genérico que aponta muitos lances que não interessam, porém todos os lances mais importantes estão contidos no resumo que atingiu cerca de 12% do tempo total da partida. Com estes resultados, o sistema traz vantagens operacionais para os seus usuários, pois é muito mais eficiente assistir cerca de 11 minutos de jogo e descartar os lances que não interessam do que assistir todos os 90 minutos de uma partida de futebol.

6.1 Próximos Passos

Apesar dos bons resultados apresentados para a sumarização de transmissões de futebol, o sistema ainda pode evoluir bastante ao aplicar futuras modificações, principalmente com o objetivo de diminuir o número de lances normais indicados erroneamente pelo sistema. Dentre as propostas, podemos citar:

- Aplicar uma análise individual dos erros de classificação para determinar possíveis fontes de problemas e, em consequência, suas respectivas soluções.

- Para reduzir a forte influência do áudio no sistema, pode-se adicionar características de vídeo que sejam mais determinantes para os bons momentos que as atuais, tais como detecção de câmera lenta, *replay*, grande área, traves, entre outros;

- O sistema atual somente classifica os lances como bom momento ou não, portanto, uma melhoria clara seria criar mais classes. Uma proposta imediata seria criar uma classe para o gol, outra para bom momento não-gol e outra para lances normais;

- Mais a frente, podem ser estudados sistemas baseados neste que contemplem a classificação de lances que possivelmente possam ser de interesse, mas que não foram abordados neste sistema, tais como cartões vermelhos e dribles que mereçam destaque;

- Melhorias no desempenho do sistema podem ser atingidas ao analisar o vídeo no domínio da compressão, pois os codificadores modernos costumam enviar informações resultantes de análises do vídeo, como, por exemplo, os vetores de movimento entre um quadro e outro de vídeo. Isto poderia substituir ou contribuir com o método desenvolvido no Capítulo 3 que aplica correlação por fase, uma operação computacionalmente custosa, para analisar o movimento de câmera durante a partida.

Apêndice A

Formulação *AdaBoost*

Pode-se entender os classificadores *AdaBoost* utilizados neste trabalho através de sua formulação original denominada como *Discrete AdaBoost* que pode ser encontrada em [96].

Sejam N amostras de um conjunto de treinamento formadas pelo par (x_n,y_n), onde x_n é uma matriz d-dimensional contendo as características a serem treinadas e y_n um vetor pertencente a $\{-1,1\}$ que representa seus respectivos rótulos.

As características e os rótulos (x_n,y_n) de cada amostra em cada iteração t do treinamento estão associados a uma distribuição de probabilidade $D_t(n)$ que corresponde à chance da amostra ser classificada como 1. A distribuição D_t é uniforme no início do treinamento e vai sendo atualizada a cada iteração.

Em cada iteração feita ao longo do treinamento, um classificador h_t é constituído com o intuito de atingir-se a melhor taxa de acerto para as características x_n. O erro de classificação em cada iteração é dado por

$$e_t = \sum_{m|h_t(x_m)\neq y_n} D_t(m) \quad , \tag{A.1}$$

onde vê-se que as probabilidades definidas em D_t são consideradas para cálculo do erro associado a classificação incorreta de determinada amostra.

De posse do erro de classificação e_t, é possível atualizar a distribuição D_{t+1} da iteração seguinte de acordo com as equações

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-e_t}{e_t}\right) \quad , \tag{A.2}$$

$$D_{t+1} = \frac{D_t(n)e^{-\alpha_t y_n h_t(x_n)}}{z_i} \quad , \tag{A.3}$$

onde z_i é um fator de normalização útil para manter sempre verdadeira a condição

$\sum_{m=1}^{N} D_t(m) = 1$.

Dessa forma, um conjunto de características x_m qualquer pode ser classificado de acordo com

$$H(x_m) = sign\left(\sum_{i=0}^{T} \alpha_i h_i(x_m)\right) \quad , \tag{A.4}$$

onde T é o número total de iterações e *sign* é o operador sinal.

A partir da definição do *Discrete AdaBoost*, é também possível definir suas variações *Real* (SCHAPIRE e SINGER [98]), *Gentle* (FRIEDMAN *et al.* [99]) e *Modest* (VEZHNEVETS e VEZHNEVETS [100]) de forma simples, pois a principal diferença entre os algoritmos é a forma como a atualização da distribuição é feita.

No classificador *Real*, os pesos necessários a atualização da distribuição D_{t+1} é atingida de acordo com

$$W_{+1} = \sum_{m|y_m h_t(x_m)=+1} D_t(m) \quad , \tag{A.5}$$

$$W_{-1} = \sum_{m|y_m h_t(x_m)=-1} D_t(m) \quad , \tag{A.6}$$

$$\alpha_t = \frac{1}{2}\ln\left(\frac{W_{+1}}{W_{-1}}\right) \quad . \tag{A.7}$$

$$\tag{A.8}$$

Além disso, no *Real AdaBoost*, diferentemente do *Discrete AdaBoost*, os classificadores retornam valores reais fazendo com que $sign(h_t(x_n))$ represente o rótulo resultante da classificação e $|h_t(x_n)|$ a confiabilidade desta classificação.

O classificador *Gentle AdaBoost* é uma evolução do *Real AdaBoost*, onde uma maior robustez e estabilidade foram atingidas através da regra de atualização dos pesos de acordo com

$$\alpha_t = \frac{1}{2}\ln\left(\frac{W_{+1} - W_{-1}}{W_{+1} + W_{-1}}\right) \quad . \tag{A.9}$$

Por fim, o *Modest AdaBoost* aborda a atualização de seus pesos de forma distinta. Chega-se a esta nova forma de atualização da seguinte maneira:

$$\overline{D}_t(m) = (1 - D_t(m))\overline{z}_i \quad , \tag{A.10}$$

$$\overline{W}_{+1} = \sum_{m|y_m h_t(x_m)=+1} \overline{D}_t(m) \quad , \tag{A.11}$$

$$\overline{W}_{-1} = \sum_{m|y_m h_t(x_m)=-1} \overline{D}_t(m) \quad , \tag{A.12}$$

$$\alpha_t = W_{+1}(1 + \overline{W}_{+1}) - W_{-1}(1 - \overline{W}_{-1}) \quad . \tag{A.13}$$

$$\tag{A.14}$$

O fator $\overline{z}_i$ também tem o objetivo de normalizar a distribuição para que esta totalize 1. Esta nova maneira de atualizar os pesos da distribuições contribuem para generalizar o classificador, mas, por outro lado, o taxa de erro provavelmente sofrerá um aumento.

Referências Bibliográficas

[1] CHANG, S.-F. "The holy grail of content-based media analysis", *Multimedia, IEEE*, v. 9, n. 2, pp. 6 –10, apr-jun 2002.

[2] SMEATON, A. F., LEHANE, B., O'CONNOR, N. E., et al. "Automatically selecting shots for action movie trailers". In: *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, MIR '06, pp. 231–238, New York, NY, USA, 2006. ACM. ISBN: 1-59593-495-2.

[3] XIONG, Z., ZHOU, X. S., TIAN, Q., et al. "Semantic retrieval of video - review of research on video retrieval in meetings, movies and broadcast news, and sports", *Signal Processing Magazine, IEEE*, v. 23, n. 2, pp. 18–27, March 2006. ISSN: 1053-5888.

[4] TJONDRONEGORO, D., CHEN, Y.-P. P., PHAM, B. "Integrating Highlights for More Complete Sports Video Summarization", *IEEE MultiMedia*, v. 11, pp. 22–37, October 2004. ISSN: 1070-986X.

[5] RUI, Y., GUPTA, A., ACERO, A. "Automatically extracting highlights for TV Baseball programs". In: *Proceedings of the eighth ACM international conference on Multimedia*, MULTIMEDIA '00, pp. 105–115, New York, NY, USA, 2000. ACM. ISBN: 1-58113-198-4.

[6] DAHYOT, R., KOKARAM, A., REA, N., et al. "Joint Audio Visual Retrieval For Tennis Broadcasts". In: *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing, Hong Kong*, pp. 561–564, 2003.

[7] NEPAL, S., SRINIVASAN, U., REYNOLDS, G. "Automatic detection of 'Goal' segments in basketball videos". In: *Proceedings of the ninth ACM international conference on Multimedia*, MULTIMEDIA '01, pp. 261–269, New York, NY, USA, 2001. ACM. ISBN: 1-58113-394-4.

[8] CHANG, Y.-L., ZENG, W., KAMEL, I., et al. "Integrated image and speech analysis for content-based video indexing". In: *Multimedia Computing and Systems, 1996., Proceedings of the Third IEEE International Conference on*, pp. 306 –313, jun. 1996.

[9] VOJKAN, M. P., PETKOVIC, M., MIHAJLOVIC, V., et al. "Multi-Modal Extraction Of Highlights From Tv Formula 1 Programs". In: *In: Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 817–820, 2002.

[10] CABASSON, R., DIVAKARAN, A. "Automatic extraction of soccer video highlights using a combination of motion and audio features." In: *Storage and Retrieval for Media Databases'03*, pp. 272–276, 2003.

[11] HANJALIC, A. "Generic approach to highlights extraction from a sport video". In: *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, v. 1, pp. I – 1–4 vol.1, 2003.

[12] DAGTAS, S., ABDEL-MOTTALEB, M. "Multimodal detection of highlights for multimedia content", *Multimedia Systems*, v. 9, pp. 586–593, 2004. ISSN: 0942-4962.

[13] REA, N., DAHYOT, R., KOKARAM, A. "Classification and representation of semantic content in broadcast tennis videos". In: *IEEE International Conference on Image Processing (ICIP'05)*, 3, pp. 1204 – 1207, Genoa, Italy, September 2005. IEEE.

[14] LEONARDI, R., MIGLIORATI, P., PRANDINI, M. "Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled Markov chains", *Circuits and Systems for Video Technology, IEEE Transactions on*, v. 14, n. 5, pp. 634 – 643, May 2004. ISSN: 1051-8215.

[15] ZHANG, D., ELLIS, D. *Detecting sound events in basketball video archive*. Relatório técnico, Dept. of Electrical Eng., Columbia University, 2001.

[16] EKIN, A. *Sports video processing for description, summarization and search*. Tese de Doutorado, Dept. of Electrical and Computer Eng., University of Rochester, 2004.

[17] COLDEFY, F., BOUTHEMY, P. "Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis". In: *Proceedings of the 12th annual ACM international conference on Multimedia*, MULTIMEDIA '04, pp. 268–271, New York, NY, USA, 2004. ACM. ISBN: 1-58113-893-8.

[18] VASCONCELOS, L., NETTO, S., BISCAINHO, L., et al. "Marcacao Automatica de Eventos Usando Sinal de Audio em Transmissoes Esportivas de TV". In: *Anais do 6o. Congresso de Engenharia de Audio AES-Brasil*, v. 1, pp. 58–64, 2008.

[19] OWENS, J. *Television Sports Production*. Focal Press, 2007.

[20] COWIE, R., DOUGLAS-COWIE, E., TSAPATSOULIS, N., et al. "Emotion recognition in human-computer interaction", *Signal Processing Magazine, IEEE*, v. 18, n. 1, pp. 32 –80, jan. 2001. ISSN: 1053-5888.

[21] GERHARD, D. *Pitch extraction and Fundamental Frequency: History and Current Techniques , November 2003*. Relatório técnico, Dept. of Computer Science, University of Regina, 2003.

[22] ROCCHESSO, D. *Introduction to Sound Processing*. Università Di Verona, 2003.

[23] TOLONEN, T., KARJALAINEN, M. "A computationally efficient multipitch analysis model", *Speech and Audio Processing, IEEE Transactions on*, v. 8, n. 6, pp. 708 –716, nov. 2000. ISSN: 1063-6676.

[24] DINIZ, P., NETTO, S., SILVA, E. D. *Digital Signal Processing: System Analysis and Design*. New York, NY, USA, Cambridge University Press, 2002. ISBN: 0521781752.

[25] SUDHIR, G., LEE, J. C. M., JAIN, A. K. "Automatic Classification of Tennis Video for High-level Content-based Retrieval". In: *Proceedings of the 1998 International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*, CAIVD '98, pp. 81–, Washington, DC, USA, 1998. IEEE Computer Society. ISBN: 0-8186-8329-5.

[26] LI, B., SEZAN, M. I. "Event detection and summarization in American football broadcast video". In: M. M. Yeung, C. S. Li, R. W. L. (Ed.), *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, v. 4676, *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pp. 202–213, December 2001.

[27] ELDIB, M. Y., ZAID, B. S. A., ZAWBAA, H. M., et al. "Soccer video summarization using enhanced logo detection". In: *Proceedings of the 16th IEEE international conference on Image processing*, ICIP'09, pp. 4289–4292, Piscataway, NJ, USA, 2009. IEEE Press. ISBN: 978-1-4244-5653-6.

[28] SEO, Y., CHOI, S., KIM, H., et al. "Where are the ball and players? Soccer game analysis with color-based tracking and image mosaick". In: Del Bimbo, A. (Ed.), *Image Analysis and Processing*, v. 1311, *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 196–203, 1997.

[29] XU, P., XIE, L., CHANG, S.-F., et al. “Algorithms and system for segmentation and structure analysis in soccer video”. In: *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, pp. 721 – 724, 2001.

[30] HUNG, M.-H., HSIEH, C.-H. “Event Detection of Broadcast Baseball Videos”, *Circuits and Systems for Video Technology, IEEE Transactions on*, v. 18, n. 12, pp. 1713 –1726, 2008. ISSN: 1051-8215.

[31] NGO, V. A., YANG, W., CAI, J. “Accurate playfield detection using Area-of-Coverage”. In: *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pp. 3441 –3444, 302010-june2 2010.

[32] SAUR, D. D., TAN, Y.-P., KULKARNI, S. R., et al. “Automated analysis and annotation of basketball video”. In: Sethi, I. K., Jain, R. C. (Eds.), *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, v. 3022, *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pp. 176–187, jan. 1997.

[33] SAUR, D. D., TAN, Y.-P., KULKARNI, S. R., et al. “Rapid estimation of camera motion from compressed video with applications to video annotation”, *IEEE Trans. Circuits Syst. Video Technol.*, v. 10, pp. 133–146, February 2000.

[34] ZHOU, W., VELLAIKAL, A., KUO, C. C. J. “Rule-based video classification system for basketball video indexing”. In: *Proceedings of the 2000 ACM workshops on Multimedia*, MULTIMEDIA ’00, pp. 213–216, New York, NY, USA, 2000. ACM. ISBN: 1-58113-311-1.

[35] LAZARESCU, M., VENKATESH, S., WEST, G. “On the automatic indexing of cricket using camera motion parameters”. In: *Multimedia and Expo, 2002. ICME ’02. Proceedings. 2002 IEEE International Conference on*, v. 1, pp. 809 – 812 vol.1, 2002.

[36] CHANG, P., HAN, M., GONG, Y. “Extract highlights from baseball game video with hidden Markov models”. In: *Image Processing. 2002. Proceedings. 2002 International Conference on*, v. 1, pp. I–609 – I–612 vol.1, 2002.

[37] LI, B., SEZAN, M. I. “Event detection and summarization in sports video”. In: *Content-Based Access of Image and Video Libraries, 2001. (CBAIVL 2001). IEEE Workshop on*, pp. 132 –138, 2001.

[38] XIE, L., XU, P., FU CHANG, S., et al. “Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models”. 2003.

[39] ASSFALG, J., BERTINI, M., DEL BIMBO, A., et al. “Soccer highlights detection and recognition using HMMs”. In: *Multimedia and Expo, 2002. ICME ’02. Proceedings. 2002 IEEE International Conference on*, v. 1, pp. 825 – 828 vol.1, 2002.

[40] KOKARAM, A., DELACOURT, P. “A new global motion estimation algorithm and its application to retrieval in sports events”. In: *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*, pp. 251 –256, 2001.

[41] PEKER, K. A., CABASSON, R., DIVAKARAN, A. “Rapid generation of sports video highlights using the MPEG-7 motion activity descriptor”. In: *Storage and Retrieval for Media Databases*, pp. 318–323, 2002.

[42] BABAGUCHI, N., KAWAI, Y., KITAHASHI, T. “Event based indexing of broadcasted sports video by intermodal collaboration”, *Multimedia, IEEE Transactions on*, v. 4, n. 1, pp. 68 –75, mar. 2002. ISSN: 1520-9210.

[43] FACON, J., TEIGÃO, R. G. N. W. “Segmentation of Soccer Video Transitions”. In: *Proceedings of 17th International Conference on Systems, Signals and Image Processing (IWSSIP 2010)*, 2010.

[44] LI, B., ERRICO, J. H., PAN, H., et al. “Bridging the semantic gap in sports video retrieval and summarization”, *J. Vis. Comun. Image Represent.*, v. 15, pp. 393–424, September 2004. ISSN: 1047-3203.

[45] REFAEY, M., ELSAYED, K., HANAFY, S., et al. “Concurrent transition and shot detection in football videos using Fuzzy Logic”. In: *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pp. 4341 –4344, 2009.

[46] BABAGUCHI, N., KAWAI, Y., YASUGI, Y., et al. “Linking live and replay scenes in broadcasted sports video”. In: *Proceedings of the 2000 ACM workshops on Multimedia*, MULTIMEDIA ’00, pp. 205–208, New York, NY, USA, 2000. ACM. ISBN: 1-58113-311-1.

[47] PAN, H., LI, B., SEZAN, M. “Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions”. In: *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, v. 4, p. IV, 2002.

[48] KAWASHIMA, T., TATEYAMA, K., IIJIMA, T., et al. “Indexing of baseball telecast for content-based video retrieval”. In: *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, v. 1, pp. 871 –874 vol.1, out. 1998.

[49] ZHONG, D., CHANG, S.-F. "Structure analysis of sports video using domain models". In: *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, pp. 713 – 716, 2001.

[50] KIJAK, E., GRAVIER, G., GROS, P., et al. "HMM based structuring of tennis videos using visual and audio cues". In: *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3 (ICME '03) - Volume 03*, ICME '03, pp. 309–312, Washington, DC, USA, 2003. IEEE Computer Society. ISBN: 0-7803-7965-9.

[51] KOBLA, V., DEMENTHON, D., DOERMANN, D. "Detection of slow-motion replay sequences for identifying sports videos". In: *Multimedia Signal Processing, 1999 IEEE 3rd Workshop on*, pp. 135 –140, 1999.

[52] KOBLA, V., DEMENTHON, D., DOERMANN, D. "Identifying Sports Videos Using Replay, Text, and Camera Motion Features". 2000.

[53] ASSFALG, J., BERTINI, M., COLOMBO, C., et al. "Semantic annotation of sports videos", *Multimedia, IEEE*, v. 9, n. 2, pp. 52 –60, 2002. ISSN: 1070-986X.

[54] PAN, H., VAN BEEK, P., SEZAN, M. "Detection of slow-motion replay segments in sports video for highlights generation". In: *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, v. 3, pp. 1649 –1652 vol.3, 2001.

[55] WANG, L., LIU, X., LIN, S., et al. "Generic slow-motion replay detection in sports video". In: *Image Processing, 2004. ICIP '04. 2004 International Conference on*, v. 3, pp. 1585 – 1588 Vol. 3, 2004.

[56] WANG, J., CHNG, E., XU, C. "Soccer replay detection using scene transition structure analysis". In: *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, v. 2, pp. ii/433 – ii/436 Vol. 2, 2005.

[57] YANG, Y., LIN, S., ZHANG, Y., et al. "Highlights extraction in soccer videos based on goal-mouth detection". In: *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, pp. 1 –4, 2007.

[58] GONG, Y., SIN, L. T., CHUAN, C. H., et al. "Automatic parsing of soccer programs". In: *Proceedings IEEE International Conference on Multimedia Computing and Systems (ICMCS)*, pp. 167–174, 1995.

[59] INTILLE, S. S., BOBICK, A. F. "Closed-world tracking". In: *Proceedings of the Fifth International Conference on Computer Vision*, ICCV '95, pp. 672–, Washington, DC, USA, 1995. IEEE Computer Society. ISBN: 0-8186-7042-8.

[60] MATSUI, K., IWASE, M., AGATA, M., et al. "Soccer image sequence computed by a virtual camera". In: *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pp. 860 –865, jun. 1998.

[61] BEBIE, T., BIERI, H. "SoccerMan-reconstructing soccer games from video sequences". In: *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, v. 1, pp. 898 –902 vol.1, out. 1998.

[62] UTSUMI, O., MIURA, K., IDE, I., et al. "An object detection method for describing soccer games from video". In: *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, v. 1, pp. 45 – 48 vol.1, 2002.

[63] INTILLE, S., BOBICK, A. "Recognizing planned, multi-person action". 2001.

[64] TOVINKERE, V., QIAN, R. J. "Detecting Semantic Events in Soccer Games: Towards A Complete Solution", *Multimedia and Expo, IEEE International Conference on*, v. 0, pp. 212, 2001.

[65] REID, I. D., ZISSERMAN, A. "Goal-directed Video Metrology". In: *Proceedings of the 4th European Conference on Computer Vision-Volume II - Volume II*, ECCV '96, pp. 647–658, London, UK, 1996. Springer-Verlag. ISBN: 3-540-61123-1.

[66] KIM, T., SEO, Y., HONG, K.-S. "Physics-based 3D position analysis of a soccer ball from monocular image sequences". In: *Computer Vision, 1998. Sixth International Conference on*, pp. 721 –726, jan. 1998.

[67] BRANCA, A., STELLA, E., ANCONA, N., et al. "Goal distance estimation in soccer game". In: *Image Analysis and Processing, 2001. Proceedings. 11th International Conference on*, pp. 565 –569, set. 2001.

[68] KANADE, T. "Eye Vision". .

[69] GUEZIEC, A. "Tracking pitches for broadcast television", *Computer*, v. 35, n. 3, pp. 38 –43, mar. 2002. ISSN: 0018-9162.

[70] PINGALI, G., JEAN, Y., CARLBOM, I. "Real time tracking for enhanced tennis broadcasts". In: *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pp. 260 –265, jun. 1998.

[71] PINGALI, G., OPALACH, A., JEAN, Y., et al. "Instantly indexed multimedia databases of real world events", *Multimedia, IEEE Transactions on*, v. 4, n. 2, pp. 269 – 282, jun. 2002. ISSN: 1520-9210.

[72] TAKI, T., HASEGAWA, J., FUKUMURA, T. "Development of motion analysis system for quantitative evaluation of teamwork in soccer games". In: *Image Processing, 1996. Proceedings., International Conference on*, v. 3, pp. 815 –818 vol.3, set. 1996.

[73] STATS, L. "SportVU". Disponível em: <`http://www.sportvu.com`>.

[74] JAIN, A. K. *Fundamentals of digital image processing*. Upper Saddle River, NJ, USA, Prentice-Hall, Inc., 1989. ISBN: 0-13-336165-9.

[75] MYLER, H. R., WEEKS, A. R. *The pocket handbook of image processing algorithms in C*. Prentice Hall PTR, 1993.

[76] GONZALEZ, R. C., WOODS, R. E. *Digital Image Processing*. 2nd ed. Boston, MA, USA, Addison-Wesley Longman Publishing Co., Inc., 2001. ISBN: 0201180758.

[77] KOKARAM, A., REA, N., DAHYOT, R., et al. "Browsing sports video: trends in sports-related indexing and retrieval work", *Signal Processing Magazine, IEEE*, v. 23, n. 2, pp. 47–58, March 2006. ISSN: 1053-5888.

[78] PEARSON, D. *Image Processing (Essex Series in Telecommunication and Information Systems)*. Mcgraw-Hill, 1991.

[79] LAO, W., HAN, J., DE WITH, P. H. N. "Automatic sports video analysis using audio clues and context knowledge". In: *Proceedings of the 24th IASTED international conference on Internet and multimedia systems and applications*, pp. 198–202, Anaheim, CA, USA, 2006. ACTA Press. ISBN: 0-88986-564-7.

[80] XIONG, Z., RADHAKRISHNAN, R., DIVAKARAN, A., et al. "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework". In: *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3 (ICME '03) - Volume 03*,

ICME '03, pp. 401–404, Washington, DC, USA, 2003. IEEE Computer Society. ISBN: 0-7803-7965-9.

[81] KIM, H.-G., ROEBER, S., SAMOUR, A., et al. "Detection of Goal Event in Soccer Videos". 2005.

[82] DENMAN, H., REA, N., KOKARAM, A. "Content-based analysis for video from snooker broadcasts", *Comput. Vis. Image Underst.*, v. 92, pp. 176–195, November 2003. ISSN: 1077-3142.

[83] DAHYOT, R., REA, N., KOKARAM, A. "Sport Video Shot Segmentation and Classification". In: *proceedings of Visual Communication and Image Processing*, Lugano, Switzerland, July 2003.

[84] ROZENN, N. R., REA, N., DAHYOT, R., et al. "Modeling High Level Structure In Sports With Motion Driven Hmms". In: *In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004*, pp. 621–624, 2004.

[85] REA, N., DAHYOT, R., KOKARAM, A. "Semantic Event Detection in Sports through Motion Understanding". In: *Proceedings of Conference on Image and Video Retrieval*, pp. 21–23, 2004.

[86] HUANG, Y.-P., CHIOU, C.-L., SANDNES, F. E. "An intelligent strategy for the automatic detection of highlights in tennis video recordings", *Expert Systems with Applications*, v. 36, n. 6, pp. 9907 – 9918, 2009. ISSN: 0957-4174.

[87] WANG, F., MA, Y.-F., ZHANG, H.-J., et al. "Dynamic Bayesian network based event detection for soccer highlight extraction". In: *Image Processing, 2004. ICIP '04. 2004 International Conference on*, v. 1, pp. 633 – 636 Vol. 1, 2004.

[88] HUANG, C.-L., SHIH, H.-C., CHAO, C.-Y. "Semantic analysis of soccer video using dynamic Bayesian network", *Multimedia, IEEE Transactions on*, v. 8, n. 4, pp. 749 –760, 2006. ISSN: 1520-9210.

[89] OKUMA, K. *Automatic acquisition of motion trajectories : tracking hockey players*. Tese de Mestrado, The University of British Columbia, Canada, 2003.

[90] OKUMA, K., TALEGHANI, A., FREITAS, N. D., et al. "A Boosted Particle Filter: Multitarget Detection and Tracking". In: *In ECCV*, pp. 28–39, 2004.

[91] MING, K., XIPENG, Q., LIDE, W. "Sports Scene Detection in TV News Video Using Variant AdaBoost Classifier", *Journal of Computer Engineering and Science*, v. 12, pp. 229–231, 2006.

[92] DUDA, R. O., HART, P. E., STORK, D. G. *Pattern Classification*. Wiley-Interscience, nov. 2001. ISBN: 0471056693.

[93] SCHAPIRE, R. E. "The Strength of Weak Learnability", *Mach. Learn.*, v. 5, pp. 197–227, July 1990. ISSN: 0885-6125.

[94] FREUND, Y. "Boosting a weak learning algorithm by majority". In: *Proceedings of the third annual workshop on Computational learning theory*, COLT '90, pp. 202–216, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN: 1-55860-146-5.

[95] KEARNS, M., VALIANT, L. "Cryptographic limitations on learning Boolean formulae and finite automata", *J. ACM*, v. 41, pp. 67–95, January 1994. ISSN: 0004-5411.

[96] SCHAPIRE, R. E., FREUND, Y. "A decision-theoretic generalization of on-line learning and an application to boosting", *J. Comput. Syst. Sci.*, v. 55, n. 1, pp. 119–139, 1997. ISSN: 0022-0000.

[97] VEZHNEVETS, A. "GML AdaBoost Matlab toolbox". `http://graphics.cs.msu.ru/en/science/research/machinelearning/`. [Online : último acesso em 2 de junho de 2011].

[98] SCHAPIRE, R. E., SINGER, Y. "Improved Boosting Algorithms Using Confidence-rated Predictions", *Mach. Learn.*, v. 37, pp. 297–336, December 1999. ISSN: 0885-6125.

[99] FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. "Additive Logistic Regression: a Statistical View of Boosting", *Annals of Statistics*, v. 28, pp. 2000, 1998.

[100] VEZHNEVETS, A., VEZHNEVETS, V. "Modest AdaBoost-teaching AdaBoost to generalize better". In: *Graphicon*, 2005.

[101] EKIN, A., TEKALP, A., MEHROTRA, R. "Automatic soccer video analysis and summarization", *Image Processing, IEEE Transactions on*, v. 12, n. 7, pp. 796–807, July 2003. ISSN: 1057-7149.