Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Conversão de Vídeo 2D para 3D em Filmagens Panorâmicas de Futebol

Maria Laura Chavez Cabrera

Documento apresentado como requisito parcial
para a conlusão do Mestrado em Informática

Orientador
Prof. Dr. Ricardo L. de Queiroz

Coorientador
Prof. Dr. Bruno Luiggi Macchiavello Espinoza

Brasília
2013

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Mestrado em Informática

Coordenador: Prof. Dr. Ricardo Pezzuol Jacobi

Banca examinadora composta por:

Prof. Dr. Ricardo L. de Queiroz (Orientador) — CIC/UnB
Prof. Dr. Adson Ferreira da Rocha — ENE/UnB
Prof. Dr. Ricardo Pezzuol Jacobi — CIC/UnB

CIP — Catalogação Internacional na Publicação

Chavez Cabrera, Maria Laura.

Conversão de Vídeo 2D para 3D em Filmagens Panorâmicas de Futebol / Maria Laura Chavez Cabrera. Brasília : UnB, 2013.
64 p. : il. ; 29,5 cm.

Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2013.

1. Visão estéreo, 2. conversão 2D para 3D, 3. segmentação,
4. reconstrução 3D, 5. vídeos de futebol.

CDU 004

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Conversão de Vídeo 2D para 3D em Filmagens Panorâmicas de Futebol

Maria Laura Chavez Cabrera

Documento apresentado como requisito parcial
para conclusão do mestrado em Mestrado em Informática

Prof. Dr. Ricardo L. de Queiroz (Orientador)
CIC/UnB

Prof. Dr. Adson Ferreira da Rocha
ENE/UnB

Prof. Dr. Ricardo Pezzuol Jacobi
CIC/UnB

Prof. Dr. Ricardo Pezzuol Jacobi
Coordenador do Mestrado em Informática

Brasília, 28 de Agosto de 2013

Dedicatória

Dedico este trabalho com todo meu carinho e meu amor para as pessoas que fizeram
tudo na vida para que eu pudesse alcançar meus sonhos,
por me motivar e segurar a minha mão
quando eu senti que o caminho estava se acabando,
a vocês para sempre meu coração
e meus agradecimentos.
Mamãe e papai

Agradecimentos

Agradeço em primeiro lugar a Deus por abençoar-me para chegar onde estou, porque fez que esse sonho acalentado se torne realidade.

Agradeço à Universidade de Brasília e ao Departamento de Ciência da Computação por me dar a oportunidade de estudar e ser um profissional.

Agradeço a meu orientador Dr. Ricardo L. de Queiroz e a meu co-orientador Dr Bruno Luiggi Macchiavello Espinoza pela orientação e ajuda que me deram para a realização desta tese, pelo apoio e amizade que me permitiu aprender muito mais que aquilo que foi estudado no projeto.

Agradeço a minha família maravilhosa que sempre me apoiou. Também agradeço aos amigos pelos conselhos, apoio, incentivo e companhia nos momentos mais difíceis da minha vida.

Resumo

A maioria dos fabricantes de TV lançaram 3DTVs no verão de 2010 usando a tecnologia de óculos de obturador. Graças a este acontecimento as aplicações de vídeo em 3D estão se tornando populares em nossa vida diária, especialmente no entretenimento doméstico. Embora cada vez mais filmes são lançados em 3D, o conteúdo de vídeo 3D ainda não é suficiente para atender a demanda do mercado. Há uma necessidade crescente de novas técnicas para converter automaticamente o conteúdo de vídeo 2D para 3D estereoscópico. Neste trabalho apresentamos um sistema para criar automaticamente vídeo estereoscópico de alta qualidade a partir de imagens monoscópicas de conteúdo esportivo, em particular, de jogos de futebol. A ideia básica de este método é separar partes estáticas e dinâmicas da cena, a parte estática é o fundo e a parte dinâmica são os jogadores. Depois exploramos informações de contexto, tais como a geometria da cena, o tamanho do jogador e o fundo conhecido. Com tais informações cria-se uma impressão de profundidade temporalmente consistente através da propagação do mapa de profundidade para uma série de quadros sequenciais pertencentes à mesma câmera, e os jogadores são modelados como cartazes. Com essa técnica se obtém resultados de vídeo 3D estereoscópico de boa qualidade.

Palavras-chave: Visão estéreo, conversão 2D para 3D, segmentação, reconstrução 3D, vídeos de futebol.

Abstract

Most TV manufacturers launched 3DTVs in the summer of 2010 using the technology of shutter glasses. Thanks to this event 3D video applications are becoming popular in our daily lives, especially in the home entertainment. Although increasingly more movies are released in 3D, the 3D video content is still not enough to satisfy the market video. There is a growing demand for new techniques to automatically convert video content from 2D to stereoscopic 3D. This work presents a system to automatically create high-quality stereoscopic video from monoscopics images of sports content, in particular, of soccer games. We take advantage contextual information, such as scene geometry, player size and the known background. With such information is created an impression of depth that is temporally consistent by propagating the depth map for a series of sequential frames belonging to the same camera and players are modeled as 'billboards'. We achieve as result stereoscopic 3D video of high quality.

Keywords: Stereo Vision, 2D to 3D conversion, segmentation, 3D reconstruction, Soccer videos.

Sumário

Lista de Figuras

Glossário

2D Bidimensional, 1

3D Tridimensional, 1

3DTV Televisões 3D, iv

DIBR Depth Image Based Rendering, em português Renderização Baseada em imagens de profundidade, 44

HSI Hue, Saturation, and Lightness, 18

LCD do inglês Liquid Crystal Display, em português tela de cristal líquido, 15

MSE Mean Squared Error, 42

NASA National Aeronautics and Space Administration, 6

RGB Red, Green and Blue, em português Vermelho Verde Azul, 18

TV Televisões, iv

Capítulo 1

Introdução

1.1 Contextualização

Vídeo é um dos principais meios de difusão de conhecimento, informação e entretenimento existentes. Apesar da boa qualidade e da boa aceitação do público, os sistemas de vídeo atuais ainda restringem o espectador a um único ponto de vista bidimensional (2D). Atualmente, alguns estudos estão sendo desenvolvidos visando oferecer ao espectador maior liberdade para decidir se ele gostaria de assistir vídeo 2D ou 3D. Desde 2010, televisões 3D (3DTV) são amplamente consideradas como uma das novas atrações, e marcas conhecidas de televisão, tais como *Sony*® e *Samsung*® lançaram televisões *3D-enabled* usando óculos de obturador 3D. Esta comercialização de 3DTV [1] é outra revolução na história da televisão depois da televisão a cores e da televisão de alta definição digital. No entanto, o sucesso na adoção de 3DTV pelo público em geral não depende apenas dos avanços tecnológicos, também dependem significativamente da disponibilidade de conteúdo de vídeo e de transmissões ao vivo em 3D, já que a criação de conteúdos estereoscópicos ainda é um processo bastante custoso e difícil, porque filmar em 3D requer *stereographers* [1] altamente treinados, plataformas estéreas caras, e um redesenho de conteúdo monoscópico. Como resultado, as técnicas para converter conteúdo 2D em 3D são uma alternativa muito importante, tanto para as novas produções como de conversão de imagens existentes. O esporte é considerado uma área que tradicionalmente está na vanguarda dos avanços tecnológicos e um excelente candidato para a visualização ao vivo em formato 3D, pois eles são extremamente populares como, por exemplo, jogos de futebol. A visão estereoscópica pode proporcionar um maior realismo ao telespectador. Por conseguinte, a produção destes programas é uma alternativa muito importante. O impacto causado pelas imagens estereoscópicas ou tridimensionais é claramente superior ao causado pelas imagens planas ou 2D. A visualização estéreo proporciona aos usuários uma ilusão de profundidade. A ilusão de profundidade em uma fotografia ou filme é criada presentando uma imagem ligeiramente deslocada para cada olho. O conteúdo 3D pode ser construído a partir de um par de imagens obtidas com duas câmeras afastados 65 milímetros uma da outra (distância média interocular humana) e estas imagens são observadas de modo que cada olho veja somente a sua imagem correspondente. Isto

[1] **o stereographer** é um profissional na área de estereoscopia e efeitos visuais que utilizam a arte e as técnicas da fotografia estéreo (fotografia 3D), ou um filme em 3D estereoscópico para criar uma percepção visual de uma imagem 3-dimensional a partir de uma superfície plana.

também pode ser obtido a partir de uma imagem e seu mapa de profundidade em tons de cinza, que contém informações sobre a distância relativa das superfícies dos objetos a partir de um ponto de vista. No entanto a geração de vídeos 3D com duas câmeras requer dispositivos especializados. Além disso, o processamento dos vídeos capturados requer *software*, *hardware* e conhecimentos especializados. Por outro lado, as câmeras monoculares são muito comuns, baratas e amplamente disponíveis. Neste trabalho propomos usar um método de renderização eficiente baseado em um vídeo monocular e de seu mapa de profundidade. O problema geral de criação de um par estéreo de alta qualidade a partir da entrada monoscópica é altamente limitada já que o processo típico de conversão a 3D consiste em estimar a profundidade para cada *pixel*, projetando cada *pixel* em um novo ponto de vista, para depois preencher os buracos que aparecem ao redor dos limites do objeto. Cada um destes passos é difícil. No caso geral necessita-se de muitos ajustes manuais dificultando a transmissão ao vivo. Em trabalhos anteriores [2] foi apresentada uma aplicação específica de conversão de 2D para 3D a partir de filmagens monoscópicas de esportes. A técnica utilizada constrói vistas panorâmicas para garantir uma profundidade estereoscópica temporalmente consistente em reconstruções de vídeo. A vista panorâmica é criada usando o mosaico de imagens [2], o mapa de profundidade é criada para a vista panorâmica gerando um modelo de fundo completo. Os mapas de profundidade para cada quadro são extraídos do panorama por projeção inversa homográfica.

Neste trabalho apresentamos um sistema para criar vídeo estereoscópico baseado em [2],onde nossa principal contribuição é uma nova técnica que gera mapas de profundidade que são temporalmente consistente em reconstrução de vídeo 3D. A consistência temporal é atingida através da propagação do mapa de profundidade, por conseguinte, a conversão de 2D para 3D é uma resposta rápida, automática, temporalmente estável e robusta. Nosso método pode ser utilizado para vistas panorâmicas, as quais são mais visualizadas nos vídeos de futebol.

1.2 Descrição do Problema

O avanço da tecnologia de video 3D permite obter um melhor beneficio econômico na industria do cinema em comparação ao vídeo 2D [3]. Embora os filmes estereoscópicos tenham tido um enorme sucesso nos cinemas, apenas recentemente o 3D em casa começou a ganhar força. A adoção de 3D ainda é limitada porque a quantidade de conteúdo 3D disponível não é suficiente e pocas transmissões ao vivo podem ser visualizados em 3D. Isto ocorre porque a criação de conteúdo estereoscópico ainda é um processo muito dispendioso e difícil. Adquirir vídeos estereoscópicos requer dispositivos especializados. Além disso, o processamento dos vídeos capturados requer *software* ou *hardware* especializado e habilidades especializadas. Por outro lado, o baixo custo das câmaras de vídeo monoculares (câmeras de vídeo comuns) faz com que estas sejam amplamente disponíveis.

A conversão de um vídeo de uma única câmera (vídeo monoscópico) para duas vistas 3D estereoscópico é um problema desafiador. Neste trabalho, propomos um método automático e eficiente de renderização para sintetizar vídeos estereoscópicos a partir de vídeos

[2]Um mosaico de imagens é uma imagem que é composto de imagens individuais menores adquiridas de diferentes pontos de vista, que foram alinhados e colados para construir uma imagem maior, dando uma visão global da cena.

monoculares de esportes (focados a vídeos de futebol). O esporte é um área que tem estado na vanguarda dos avanços tecnológicos em televisão, como por exemplo na difusão da televisão em alta definição (HDTV). Esportes são excelentes candidatos para a visão estereoscópica ao vivo, por serem extremamente populares, e poderem se beneficiar do aumento do realismo que proporciona a visualização estereoscópica. O método proposto tira vantagem a informações de contexto, tais como a geometria do campo, o tamanho de jogador e o fundo conhecido. Nós criamos uma impressão de profundidade temporalmente consistente através da propagação do mapa de profundidade para uma série de quadros sequenciais pertencentes à mesma câmera, e os jogadores são modelados como cartazes ou *billboards*. O método proposto pode ser usado em sequências de vídeo com vistas panorâmicas, já que as vistas panorâmicas predominam o tempo de visualização dos vídeos de futebol.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo do trabalho é aplicar novas tecnologias de vídeo 3D e desenvolver um sistema que possa criar automaticamente vídeo estereoscópico de alta qualidade a partir de filmagem monoscópica, especificamente para eventos esportivos que são filmados com o movimento de câmera panorâmica [4].

1.3.2 Objetivos Específicos

- Implementar a técnica de segmentação por cor para identificar esportistas e campo de jogo (fundo) dentro de uma sinal de vídeo digital de conteúdo esportivo.
- Implementar a técnica de estimação de profundidade evitando a cintilação (*flickering*).

1.4 Estrutura do Trabalho

A presente dissertação é construída de acordo com os objetivos propostos. Este documento apresenta descrição dos conceitos, revisão bibliográfica e técnica utilizada para o desenvolvimento do sistema, além da descrição do projeto e dos resultados alcançados. A monografia está organizada em 5 capítulos contando com esta introdução.

O Capítulo 2 caracteriza o estado-da-arte. Aqui falamos da estereoscopia, o formato 2D + profundidade, segmentação de imagens e operações de pós-processamento, e finalmente apresentamos a revisão bibliográfica em relação a conversão de vídeo monoscópico para vídeo estereoscópico.

O Capítulo 3 tem como finalidade a descrição da metodologia que propomos para a conversão de vídeo 2D para 3D em filmagens panorâmicas de futebol.

No Capitulo 4 são analisados e discutidos os resultados provenientes do desenvolvimento do método proposto.

Por último, no Capitulo 5, apresentam-se as conclusões oriundas de todo o projeto de investigação.

Capítulo 2

Fundamentação Teórica

2.1 Estereoscopia

O objetivo básico da estereoscopia, que remonta aos primeiros tempos da arte fotográfica, é obter uma imagem de relevo, ou seja, introduzir uma sensação tridimensional de forma que possamos perceber profundidade. Ao longo da história foram desenvolvidas diferentes técnicas para alcançar essa sensação de 3D.

2.1.1 Definição Estereoscópica

Ao olharmos para uma cena qualquer, nosso cérebro esta calculando o tamanho, distância, posição, profundidade dos objetos presentes no ambiente, através das imagens captadas pelos olhos. O cérebro, então, cria a sensação da visão 3D, esta simulação é chamada estereoscópica [5].

A palavra *estereoscopia* deriva do grego *stereo* que significa *solido ou relevo*, e *skopein* que significa *ver ou olhar*. Assim, estereoscopia pode ser entendida como *visão em relevo*. A frequente interpretação de *estéreo* no sentido de *dois* é resultante do fato de necessitarmos de dois olhos e dois ouvidos para vermos e ouvirmos especialmente.

O termo estereoscopia engloba todas as técnicas que utilizam o mecanismo visual binocular para criar a sensação de profundidade em duas ou mais imagens. A profundidade é simulada apresentando-se duas imagens bidimensionais do mesmo objeto ligeiramente diferentes para cada olho (Figura 2.1), e através de técnicas de visualização, que podem ser a olho nu ou com uso de óculos especiais, o efeito 3D é criado.

2.1.2 História da Estereoscopia

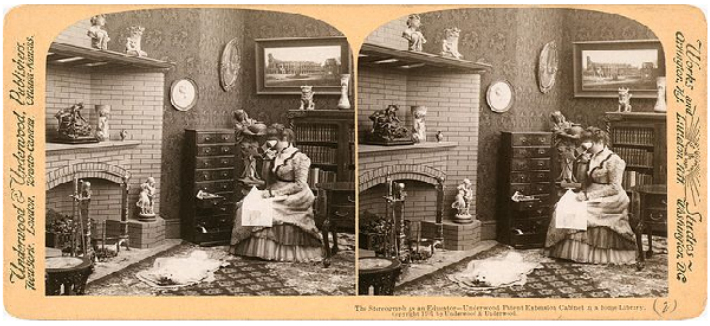
Retomando os arquivos da história, vemos que a primeira tentativa de aprender o funcionamento da visão estéreo foi na época do Renascimento, quando Euclides, Galiano e Leonardo da Vinci observaram e estudaram o fenômeno da visão binocular, por esta razão, foram considerados os pioneiros nesta área. O astrônomo Johan Kepler também realizou estudos sobre a estereoscopia, que precedeu à fotografia.

Em 1832, Charles Wheatstone, físico e professor de filosofia do Málico King's College de Londres, inventou o estereoscópio, recriando o fenômeno natural da visão binocular usando um aparelho com espelhos [7]. Por este dispositivo podia-se ver desenhos geométricos em

Figura 2.1: Arquivo Fotográfico e Digital da Biblioteca Nacional de Chile[1]

relevo. Ele apresentou um artigo na "Royal Society" de Londres em 1833 intitulado "On some remarkable and hitherto unobserved phenomena of binocular vision". Já em 1838 Wheatstone inventou os estereogramas. Um estereograma é um par de imagens que diferem em deslocamento lateral de elementos.

Anos mais tarde, em 1849, Sir David Brewster projetou e construiu a primeira câmera fotográfica estereoscópica, com este se obteve as primeiras fotografias em relevo. Construiu também um visualizador com óculos para observar as fotografias. Mais tarde, Oliver Wendell Holmes, em 1862, construiu outro modelo estereoscópico de mão que se tornou muito popular no final do século XIX [8], com o qual podia-se ver em relevo fotografias estereoscópicas montadas num cartão. A Figura 2.2 mostra uma fotografia estereoscópica antiga junto com o visor de Holmes para vê-lo em relevo. Extensas coleções foram criados e colocado à venda. Podiam-se achar fotografias em relevo em qualquer lugar do mundo.



(a) Rainha Vitoría [2]

(b) Visor de tipo Holmes©[3]

Figura 2.2: (a) Amostra de fotografia estereoscópica montada num cartão, para vê-lo em relevo era necessário um visor do tipo Holmes(b).

Durante os anos 1930, houve um ressurgimento da estereofotografia e, como consequência, houve o aparecimento de câmeras 3D com películas de 35 milímetros como a *Realist*[9] ou a *ViewMaster*, o que facilitou a obtenção de tais imagens. Essas câmeras não são mais fabricadas, e hoje são objetos de coleção.

[1]Fonte: http://www.rebootarobot.com/estereoscopia

[2]Fonte: http://www.loc.gov/pictures/resource/ppmsca.08781/ Acessado em Março 2012

[3]Fonte: http://en.wikipedia.org/wiki/File:Holmes_stereoscope.jpg Acessado em Março 2012

(a) Câmera estéreo *Realist*©[4] (b) Modelo G da *ViewMaster*©[5]

Figura 2.3: (a) Câmera estéreo *Realist*, (b) Modelo G da *ViewMaster*.

Na década de 1950 apareceram os primeiros filmes em 3D e tentou-se a exploração comercial destes, mas com pouco impacto no mercado cinematográfico. Alem disso, alguns filmes que foram feitos tinham problemas de visão, causados pela falta de conhecimento dos técnicos da época sobre a problemática que traz um filme estereoscópico, o que causou desconforto visual fazendo que parte do público rejeitasse este tipo de cinema.

Na década de 1980, se alcançaram resultados mais espetaculares, com sistemas de grande formato do cinema como o *IMAX* [12], para obter imagens de alta resolução em telas gigantes, depois de grandes investimentos em pesquisa.

Na década de 1990 o progresso da informática permitiu apresentar imagens em 3D em monitores de computador. Os computadores também permitem gerar imagens espetaculares de síntese em relevo para aplicações científicas, industriais ou de entretenimento.

Durante a primeira década do novo milênio retomou-se o interesse em recriar as imagens 3D estereoscópicas com um alto grau de definição. Por isso, as câmeras de vídeo estereoscópico são fabricados principalmente para uso em cinemas e televisão. Recentemente a *NASA* tem usado a estereoscopia como uma ferramenta para visualizar e analisar as imagens em 3D de Marte enviadas pela sonda *Pathfinder* [13], a fim de obter dados mais realistas.

2.1.3 Princípios Biológicos

O cérebro é capaz de criar a sensação de profundidade, devido à estereoscopia, também chamado visão estéreo ou binocular. O cérebro tem a habilidade de fazer a fusão das imagens horizontalmente díspares, que são captadas por cada olho. Esta disparidade ocorre porque os olhos estão separados no plano horizontal da cabeça, tendo assim visões diferentes da cena.

A (Figura 2.4) apresenta a geometria envolvida no processo de percepção de profundidade. Neste esquema os dois olhos são apresentados por duas lentes simples com centros localizado em C_1 e C_2. Os pontos O_1 e O_2 correspondem ás origens das duas retinas (por simplicidade as retinas estão representadas em frente dos olhos). Para cada olho a luz originada do ponto P é agrupada pela lente e focalizada em apenas um ponto na

[4]Fonte: http://www.3dham.com/stereo_cameras/index.html Acessado em Julho, 2013

[5]Fonte: http://www.guardian.co.uk/artanddesign/2008/jul/31/viewmaster.design.classic Acessado em Julho, 2013

retina. Quando um raio de luz que passa pelo centro da lente, ele não muda da direção. Podemos então determinar a posição na qual o ponto é projetado. Nesse caso podemos simplesmente conectar o ponto P com C_1 e C_2 com duas linhas retas. A intersecção destas linhas com a retina são os pontos projetados P_1 e P_2. A imagem nos mostra que as posições destes pontos diferem das origens de suas retinas. As coordenadas desses pontos são respectivamente $X_1 = P_1 - O_1$ e $X_2 = P_2 - O_2$. Assim, a disparidade das retinas é definida por $X_2 - X_1$. Através da triangulação, podemos deduzir que a distância perpendicular do ponto P, ou seja, sua profundidade é inversamente proporcional à disparidade. A medida que P se desloca para uma distância infinita a disparidade diminui. O cérebro utiliza a disparidade das imagens para determinar a profundidade dos objetos e, assim, a percepção 3D. Por isso, é difícil definirmos a distância de objetos quando observamos com apenas um olho.

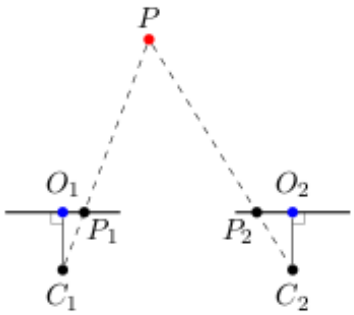


Figura 2.4: Exemplo geométrico de como nosso cérebro é capaz de determinar a distância de um objeto [14].

A terceira dimensão é capaz de reconstruir-se em nosso cérebro através de complexos processos fisiológicos e psicológicos relacionadas com a visão monocular e binocular.

- **Visão Monocular:** Quando olhamos com um olho só, criamos uma imagem plana 2D, mas com informação intuitiva de profundidade e de distância.
- **Visão Binocular:** Quando olhamos com os dois olhos, este fornece maior informação espacial para a terceira dimensão permitindo a sensação de volume.

2.1.4 Mecanismos Intuitivos da Visão Binocular

A disparidade da visão não é a única informação sobre o ambiente tridimensional. Existem outras informações que auxiliam na percepção de profundidade pelo cérebro [14].

Distribuição de Luz e Sombra

A distribuição de luz e sombra pode criar a ilusão de volume. A iluminação é um fator intuitivo de volume muito importante, pois a sombra e o contraste fornecem maior sensação do relevo e de volume. Por exemplo (Figura 2.5), um círculo pintado pode ser convertido numa esfera somente com sombrear e escurecê-lo simulando iluminação. Este é uma das técnicas potencias utilizados por programas de renderização para criar ilusão 3D. Isto permite prover sensação de profundidade num monitor 2D simples.

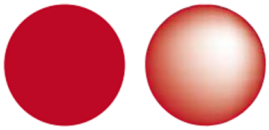


Figura 2.5: Distribuição de Luz e Sombra©[6]

Superposição de Imagens

Objetos que estão relativamente mais próximos ao observador normalmente ocultam ou sobrepõem objetos mais distantes. Por exemplo na Figura 2.6 as duas imagens deveriam representar exatamente a mesma coisa com a diferença de que em uma (primeira, à esquerda) se vê o elemento nuvem completa e parcialmente escondido a lua. Assim, a sensação que transmite ao nosso cérebro é a de nuvem está diante da lua. No outro (à direita) acontece o contrário, a lua está completamente e parece que está em frente da nuvem.

Figura 2.6: Superposição de Imagens ©[6]

Perspectiva

Linhas fisicamente paralelas convergem para um ponto comum no horizonte, como por exemplo, linhas de trem. Por exemplo, na Figura 2.7 as árvores são exatamente do mesmo tamanho, mas uma está mais próxima ao ponto de fuga e, portanto, parece estar mais distante. Para compensar esta contradição de que estaria mais longe, mas é de igual tamanho, chegamos a pensar que o objeto mais distante é do mesmo tamanho na imagem, porque seria maior do que o outro.

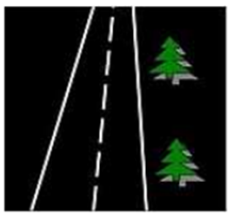


Figura 2.7: Perspectiva ©[6]

[6]Fonte: http://commons.wikimedia.org/wiki/Category:Stereoscopy Acessado em Julho, 2013

Tamanho Aparente

Objetos que aparecem pequenos na retina são interpretados como distantes e objetos grandes, como próximos. A medida que se aproximam, os objetos parecem maiores. No exemplo Figura 2.8, vemos como nosso sistema visual cria a sensação de que o objeto A é maior que B, a reconstrução espacial se faz comparando as sensações visuais e em função deles interpreta as distâncias para os objetos encontrados. Assim, o cérebro interpreta que o objeto A está mais perto do B porque ele é percebido em um tamanho maior, e vice-versa.

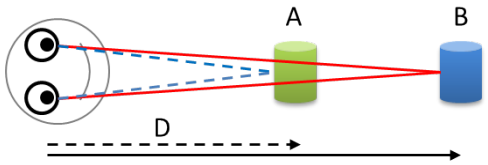


Figura 2.8: Tamanho aparente ©[6]

Movimento Paralaxe

A paralaxe é definida como o deslocamento aparente da posição de um objeto causada pela mudança do ponto de observação, tendo como referencia uma linha ou a um ponto. A paralaxe é medida ao longo do eixo X. A Paralaxe é tanto maior quanto mais próximo estiver o objeto do observador móvel. Por exemplo, faça uma experiência como na Figura 2.9, levante um dedo e, com um olho fechado, alinhe seu dedo com um livro ou objetos sobre uma mesa. Agora, olhe para o dedo com o outro olho. Ele parecerá estar numa posição diferente em relação ao segundo plano. O fundo, porém, não parece sofrer esse "deslocamento". O aparente movimento varia em função da distância entre o dedo e o olho. Quanto mais próximo, mais o dedo parecerá se mover.

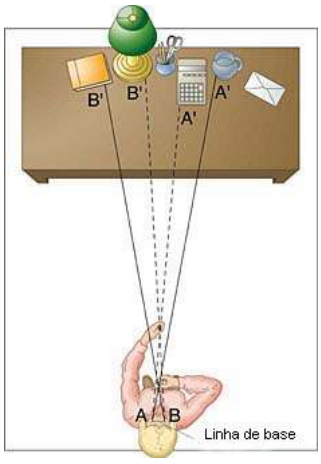


Figura 2.9: Movimento paralaxe©[7]

[7]Fonte: http://www.geocities.ws/saladefisica5/leituras/paralaxe.html Acessado em Agosto, 2013

2.1.5 Métodos de gravação de imagens 3D

Tanto o formato de exibição simultânea (lado-a-lado), quanto o sequencial são utilizados nos monitores e televisões 3D. Com o auxilio de óculos, ativos ou passivos, é possível a visualização das imagens em três dimensões. Existem também tecnologias em que não é necessário o uso de óculos para ver o conteúdo 3D. Os monitores e TVs são capazes de receber as sinais 3D e recodificá-los em tempo real para seu próprio formato e especificação que permitem a exibição das imagens por este aparelho [15].

Sequencial

Neste método os quadros para o olho esquerdo e direito são alternados, como mostra a (Figura 2.10). O formato de *frame* sequencial é bastante utilizado, fazendo parte das especificações do padrão *Blu-Ray 3D* [15]. Além disso, o método sequencial é ideal para a visualização com a tecnologia obturador ativo, que consiste na sincronização dos quadros para cada olho com os óculos ativos que abrem e fecham de acordo com a imagem que esta sendo exibida na tela. Existem também óculos passivos onde as lentes possuem filtros, deixando passar somente a luz emitida pela imagem esquerda ou direita para cada olho respectivamente. Esta tecnologia será detalhada em seções posteriores. Como ponto negativo, neste tipo de exibição ainda é possível notar uma oscilação da imagem, devido à troca de imagem de um olho para ou outro.

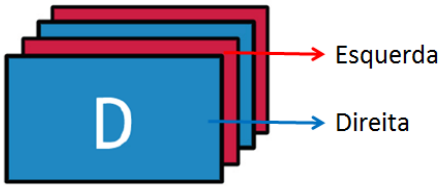


Figura 2.10: Exibição de quadros sequenciais 3D. Quadros com imagens para cada um dos olhos que são alternados. ©[8]

Lado-a-lado

No formato de 3D lado-a-lado, um quadro é dividido ao meio horizontalmente e cada uma das imagens para o olho esquerdo e direito ocupam metade deste quadro, como mostra a (Figura 2.11)

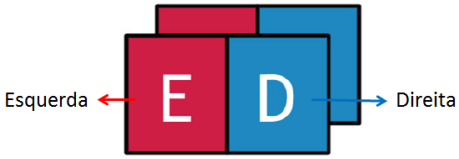


Figura 2.11: Na exibição de 3D lado-a-lado cada quadro contém a representação da imagem para o olho esquerdo e direito ©[8].

[8]Fonte: http://www.jvc.eu/3d_monitor/technology/video.html Acessado em Agosto, 2013

O quadro que corresponde ao olho esquerdo, é reduzido para a metade esquerda da imagem e o correspondente ao olho direito, é reduzido para a metade direita da imagem. Para um quadro com resolução de 720p, ou seja com resolução 1280 x 720 *pixels*, cada uma das imagens teria uma resolução de 640 x 720 *pixels*.

Apesar de ser gravado com este método, o conteúdo normalmente não é exibido lado a lado pelos monitores 3D, e sim na forma sequencial. Para isso é feita uma conversão. O quadro recebido é divido e cada imagem é novamente aumentada para sua resolução original através de algoritmos de escalonamento. Então, elas são enviadas de maneira alternada, para adequar-se ao formato 3D sequencial, como mostra a (Figura 2.12). O motivo da conversão é exitar o treinamento necessário para visualizar uma imagem 3D no formato Lado-a-lado.

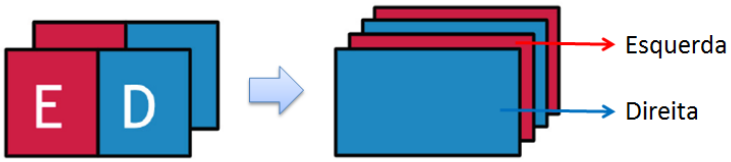


Figura 2.12: Conversão de 3D lado-a-lado para sequencial ©[8].

Esquerda direita independente (*LR Independent*)

Nesse formato os quadros da esquerda e da direita possuem a resolução máxima. Diferente do formato lado-a-lado em que cada imagem possui a metade da resolução. A imagem transmitida possui o dobro da resolução horizontal. Os padrões de codificação MPEG2 e H.264 estão estabelecidos para este formato [16]. O sistema desenvolvido neste trabalho gera e transmite o conteúdo lado-a-lado esquerda e direita independente (Figura 2.13).

Figura 2.13: Formato 3D *LR Independent* ©[8]

2.1.6 Métodos de exibição de imagens 3D

Na maior parte das tecnologias de monitores e televisões 3D atuais, o uso de óculos especiais é necessário para a percepção de profundidade e para a visão estereoscópica. Como descrito anteriormente, os métodos de exibição em 3D mostram imagens separadas da mesma cena com conteúdo adequado para o olho esquerdo e direito. O princípio básico do óculos 3D é permitir que a imagem certa chegue somente para o olho correspondente e seja bloqueada para o outro. Os dois tipos de óculos 3D são ativo e passivo.

Óculos passivos

A separação das imagens ocorre pelo tipo de cada uma das lentes. Existem dois tipos de óculos passivos: anáglifos e polarizados. O funcionamento destes óculos é descrito a seguir.

- **Óculos anáglifos:** os óculos de duas cores (anáglifos) são o tipo mais simples de óculos passivos para visualização em 3D. Esta técnica foi desenvolvida pelo alemão Wilhelm Rollmann, em Leipzig, no ano de 1853. As cores mais comuns para as lentes são o azul e vermelho (Figura 2.14).Também são utilizadas as cores verde/vermelho e amarelo/azul. Podem ser utilizadas quaisquer cores complementares [17].

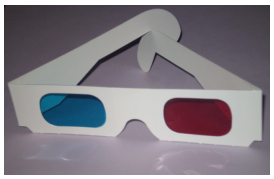


Figura 2.14: Óculos anáglifos coloridos com lentes azul e vermelha.

Nesse tipo de óculos, as imagens para o olho esquerdo e direito são projetadas na mesma tela, cada uma com uma cor diferente, ou impressas na mesma fotograma, e devido a cor da tinta de cada lente, cada olho recebe uma imagem separada, como visto na Figura 2.15.

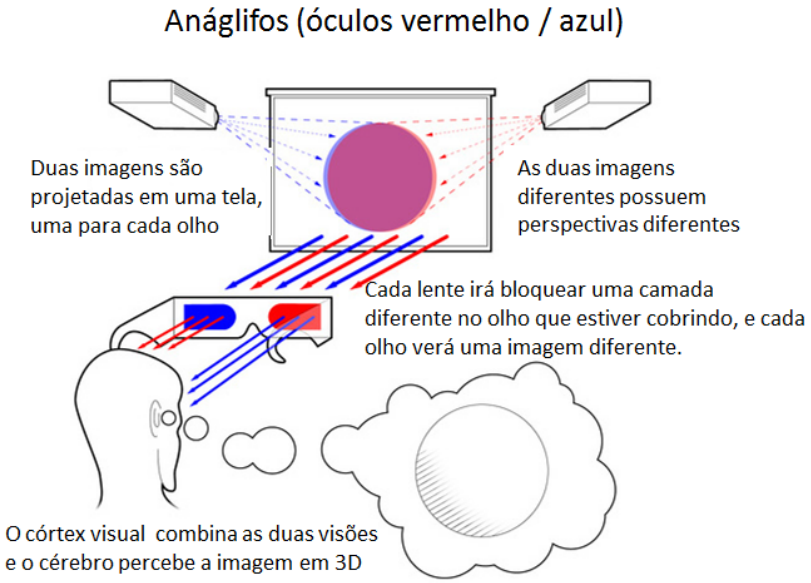


Figura 2.15: Funcionamento de óculos anáglifos coloridos ©[9]

A vantagem deste método é o baixo custo, pois não é necessário nenhum hardware adicional para os monitores e impressoras e os óculos custam centavos, podendo até

[9]Fonte: http://www.onlineschools.org/blog/how-3d-works/ Acessado em Março, 2012

ser feitos em casa. Apesar disto, esta tecnologia possui muitas limitações, como, má qualidade da imagem, provendo uma simulação pobre da realidade podendo até causar mal estar nas pessoas.

- **Óculos de lentes polarizadas:** Atualmente, a tecnologia de óculos 3D passivos que vem sendo utilizada é a polarização. Este método é utilizado para a exibição de filmes comerciais nos cinemas em 3D. A polarização faz com que a luz não se propague para todos os lados, permitindo, assim, que seja desviada e refletida para a direita ou para a esquerda. Cada lente desse tipo de óculos 3D possui uma polarização oposta e na tela são projetadas duas imagens com as respectivas polarizações para que o olho esquerdo e direito recebam as imagens separadamente.

  Nesta tecnologia, duas imagens são projetadas na mesma tela com raios de luz em diferentes direções. Quando a luz encontra uma lente que está polarizada na direção oposta, ela não consegue passar. Por outro lado, se a lente e a luz possuem a mesma polarização a luz pode atravessar e chegar ao olho.

  Há dois tipos de polarização, que podem ser simulados com uma mola *slink*:

  A polarização plana ou linear: (Figura 2.16) é produzida quando oscilamos uma das extremidades da corda apenas na vertical (ou na horizontal). O espectador utiliza óculos polarizados linearmente que contem um par de filtros polarizados ortogonais orientada na mesma direção que o projetor. Cada um dos filtros deixa passar apenas a luz que é polarizado na mesma direção e bloqueia a luz polarizada ortogonalmente, cada olho vê somente uma das imagens projetadas, obtendo-se o efeito 3D. Óculos de polarização linear exigem que o espectador mantenha a cabeça na mesma direção que a luz polarizada, caso contrario o efeito 3D poderia ser perdido.

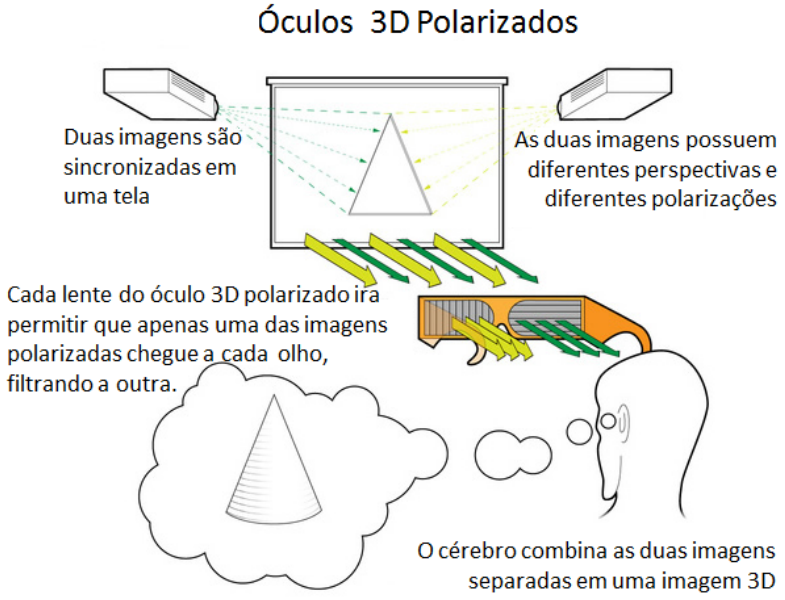


Figura 2.16: Funcionamento de Óculos Anáglifos Polarizados ©[9].

A polarização circular: produzida quando oscilamos a corda com velocidade constante ao longo de uma circunferência [19]. Nessa situação, pode haver dois tipos de ondas resultantes, dependendo do sentido de rotação, horário ou anti-horário. A projeção envolve duas imagens, projetadas com feixes de luz que possuem polarizações opostas, por exemplo, a imagem que será vista pelo olho esquerdo é projetada com polarização horária (Figura 2.17-a), e a imagem vista pelo olho direito tem polarização anti-horária (Figura 2.17-a). O resultado é semelhante ao de visualização estereoscópica usando óculos polarizados linearmente, exceto que o espectador pode inclinar sua cabeça e ainda manter a separação esquerda/direita

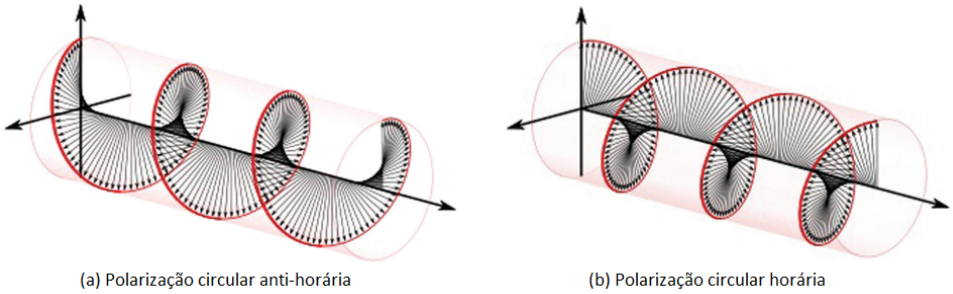


Figura 2.17: (a) Polarização circular anti-horária (b)Polarização circular horária ©[20]

Algumas vantagens dos óculos 3D polarizados são: ótima separação da imagem para cada olho, a visualização das imagens não possui oscilações e possibilita a troca entre a visualização 2D e 3D pois os óculos não modificam a imagem visualizada.

Algumas desvantagens são: a tecnologia de projeção polarizada é muito cara e é difícil de manter a polarização da luz em uma tela de TV ou monitor de computador. Para utilização desta tecnologia nos cinemas é necessária uma tela especial que reflete a luz para manter a polarização das imagens projetadas. Na TV o problema é de colocar os filtros nos *pixels* certos.

Óculos ativos

Óculos com obturador ativo (*active shutter*) ilustrado na Figura 2.18, estão sendo bastante utilizados com os aparelhos de tecnologia 3D, pois para o seu uso não são necessárias grandes modificações nos monitores de alta definição existentes.

Figura 2.18: Óculos *Active Shutter* NVidia 3D ©[10]

[10]Retirado de: http://www.nvidia.com/object/3d-vision-main.html Acessado em Março, 2012

As lentes desses óculos tem uma camada de cristal líquido. Quando uma voltagem é aplicada, a lente fica escura e não permite que a luz atravesse. As lentes da esquerda e da direita são fechadas de forma alternada em sincronia com a taxa de atualização da tela. O óculos é conectado ao monitor através de *bluetooth*, infra-vermelho ou rádio. Os quadros para o olho esquerdo e direito são mostrados alternadamente, e os óculos garante que cada olho vai ver apenas as imagens certas. Este processo é ilustrado na Figura 2.19.

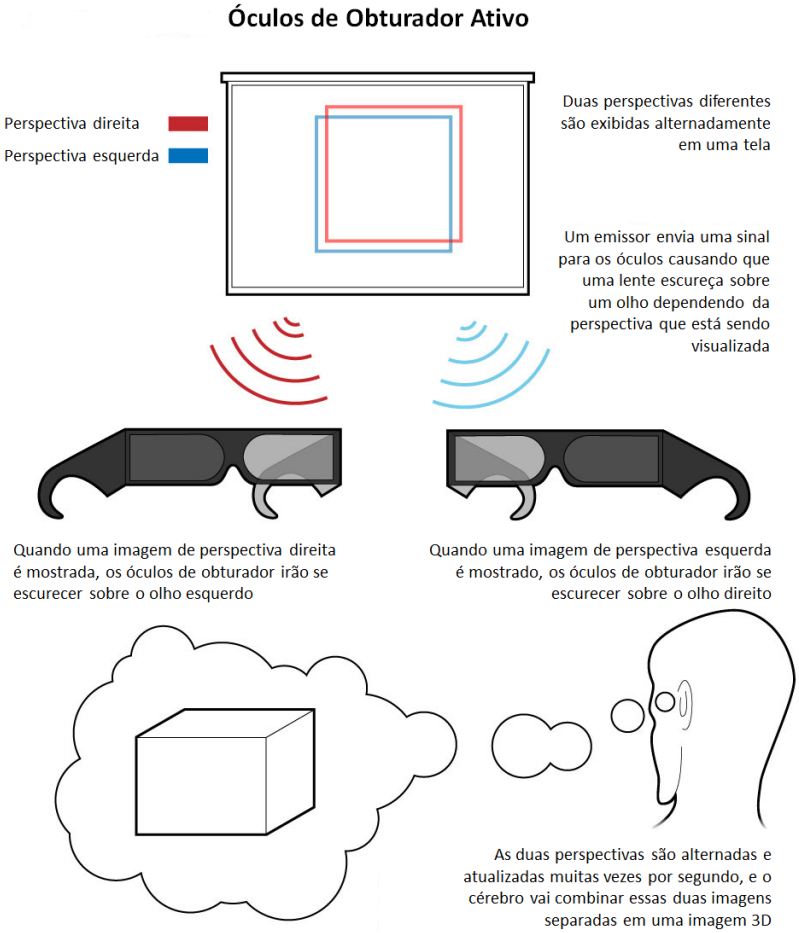


Figura 2.19: Funcionamento de Óculos *Active Shutter*©[18].

Esta tecnologia é a mais madura em termos de 3D para uso não-profissional. Além disso, a experiência 3D é a melhor quando comparada com os óculos passivos.

Esta tecnologia pode chegar mais facilmente à casa do usuário comum, já que pode ser utilizada nos monitores LCD existentes de plasma. As únicas modificações necessárias seriam o aumento da taxa de atualização das imagens e o mecanismo de sincronização entre os óculos e o conteúdo mostrado. Estes óculos propiciam uma ótima qualidade da imagem.

Este tipo de óculos também podem causar oscilações na imagem, principalmente nas telas com taxa de atualização de 120 Hz. Além disto os óculos são alimentados por baterias devido a necessidade de energia para "fechar" as lentes. Outra dificuldade é a comunicação e sincronização entre os óculos e o monitor ou televisão.

2.2 2D + profundidade

É uma alternativa para o vídeo estéreo convencional, no qual as imagens estéreos são geradas por interpolação do vídeo e de dados de profundidade da imagem.

Este formato fornece uma melhor eficiência de compressão. O intervalo de profundidade é limitado a um intervalo entre dois pontos de *Z perto* e *Z longe* indicando a distância máxima e mínima de um ponto 3D à câmera. O intervalo de profundidade é quantificado linearmente com 8 *bits*. Por exemplo, os pontos mais próximos recebem o valor 255 e os mais distantes o valor 0, obtendo um mapa de profundidade representada por uma escala de cinza. Os mapas de profundidade podem ser obtidos por sensores ou gerados por algoritmos computacionais.

Figura 2.20: Codificação de uma imagem: *2D plus Depth Map* [21].

2.2.1 Geração estéreo de 2D + profundidade

O processo de conversão estereoscópico para imagens pode ser dividido em :

- Análise de vídeo ou imagem.
- Processo de estimação de profundidade.
- processo de geração de paralaxe.
- processo de síntese estereoscópica.

como é mostrado na Figura 2.21

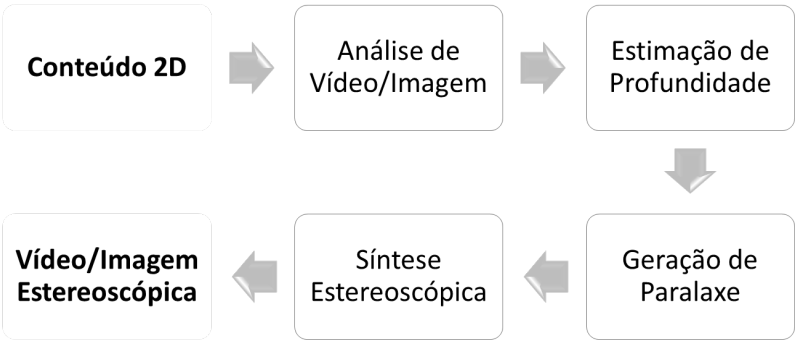


Figura 2.21: O processo de conversão de 2D para estereoscopia.

Geralmente, na fase de análise de vídeo ou imagem, a imagem é segmentada e são determinadas critérios para estimar a informação de profundidade [22]. O processo de segmentação de imagem é classificado em métodos manuais, semi-automático e automático. A segmentação manual de imagens é realizado manualmente pelos usuários com ferramentas extras adicionais. Em caso de segmentação automática de imagem, a imagem é dividida automaticamente usando a cor, borda ou informações adicionais extraídas da imagem. Se a imagem contém texturas complexas, a segmentação automática pode não produzir o resultado desejado. Assim, após a realização da segmentação automática é comum que, o usuário realize adicionalmente segmentação manual.

O processo mais difícil na conversão estereoscópica é a fase de extração de informação de profundidade em uma imagem. Na etapa de análise de vídeo/imagem se determina algumas sugestões para extrair informação detalhada e com essa informação poder gerar o mapa de profundidade . Em [22], o tipo de imagem é determinada através da classificação da imagem, e em seguida, de acordo com o tipo de imagem, um valor diferente de profundidade é atribuído a cada região de segmentação.

2.3 Segmentação

A segmentação de imagens digitais [23] é uma parte importante de muitas tarefas na análise e processamento digital. Atualmente existe um grande número de trabalhos sobre os diferentes técnicas, modelos e algoritmos para segmentar imagens a cores.

A segmentação de imagem refere-se a partição de uma imagem em regiões que são homogêneas em relação a algum aspecto da imagem. A segmentação de imagem é geralmente a primeira tarefa no processo de análise de qualquer imagem. As tarefas que acompanham a segmentação, tais como, extração de características ou reconhecimento de objetos, dependem fortemente de como encontrar a imagem segmentada depois de uma etapa de pré-processamento, que facilita a segmentação. Um bom algoritmo de segmentação é fundamental para o reconhecimento de objetos. Por outro lado, uma super-segmentação de imagem pode dividir um objeto em muitas regiões diferentes, enquanto sub-segmentar a imagem pode juntar vários objetos em uma região. Portanto, a etapa de segmentação será um fator determinante para o sucesso ou o fracasso da análise. As técnicas de segmentação de imagens se dividem em quatro grupos:

- Segmentação baseada no valor de *pixel*.

- Segmentação baseada em regiões.
- Segmentação usando a detecção de bordas.
- Segmentação baseada na física.

2.3.1 Segmentação Baseada no Valor de *Pixel*

A segmentação baseada no valor de pixel inclui algumas técnicas baseadas no histograma. Trata-se de obter o histograma de uma imagem, na qual um ou mais picos são identificados, e os intervalos que os rodeiam são utilizados de imediato no processo de classificação de pixel. Dentro deste grupo são consideradas: a segmentação por agrupamento de *pixels* em um espaço de cor e, de acordo com a característica, os algoritmos de agrupamento difusos e limiarização.

Espaço de Cor

O espaço de cores ou sistemas de cores são especificações de coordenadas onde cada ponto deste corresponde a uma cor [24], permitindo representar os *pixels* de uma imagem digital. Há muitos espaços de cor, como o espaço RGB e HSI que são a representação mais comum de imagens digitais.

O espaço de cores RGB é o espaço de cor mais difundido pela grande maioria dos dispositivos comuns, tais como as câmeras de vídeo e fotográficas. Este espaço é caracterizado por um cubo de lado R, comprimento G e altura B, tendo sido idealizado por Thomas Young (1773-1829). Neste cubo (Figura 2.22) três dos vértices são ocupados pelas cores Vermelha (R), Verde (G) e Azul (B). O modelo de espaço de cores RGB é representados por 8 *bits*, neste caso, tem-se um número total de cores de $(2^8)^3$, ou seja, mais de 16 milhões de cores. A cor preta é definida na origem do sistema e a cor branca é dada pelos três valores máximos do espaço RGB definido. Além disso, a gama cromática em tons de cinza é representado pela diagonal do cubo.

Imagem Digital

Uma imagem digital é caracterizada por uma função em duas dimensões $f(x, y)$, em que x e y são as coordenadas de um determinado ponto, denominado de *pixel* (*picture element*), e a amplitude de f de qualquer par ordenado (x, y) é chamado de intensidade ou nível de cinza. Esta função pode ser representada matematicamente por uma matriz, em que cada elemento desta matriz é a amplitude de f e as coordenadas x e y correspondem às colunas e linhas, respectivamente [24].

As imagens digitais podem ser coloridas ou em níveis de cinza. No sistema RGB, por exemplo, as imagens coloridas assumem três canais que são os canais R (Red), G (Green) e B (Blue), sendo representada por uma matriz para cada canal [24]. Pode-se denominar de vídeo digital uma sequência de imagens digitais e que estas são denominadas de *frames*.

Segmentação por Cor

Segmentação é um processo que divide uma imagem em regiões. A partir de um conjunto de amostras representativo das cores de interesse, pode-se obter uma estimativa

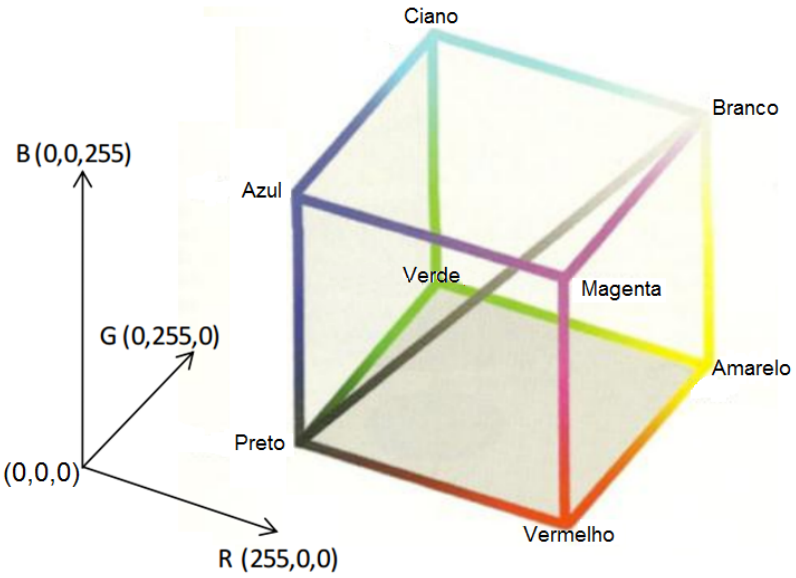


Figura 2.22: Cubo RGB [25].

da cor média que se deseja segmentar na imagem. O processo de segmentação consiste em classificar cada *pixel* da imagem dentro da faixa de interesse especificada para a cor. Para realizar esta classificação, utiliza-se a limiarização, um dos algoritmos de segmentação mais intuitivos e rápidos empregados em processamento de imagens [24].

A limiarização funciona através do estabelecimento de um limiar T que separa a imagem em duas regiões. Os *pixels* da imagem cujos valores são maiores do que T são chamados de pontos do objeto e os pontos da imagem cujos valores sejam inferiores a T são chamados de fundo. Dessa forma, pode-se criar uma imagem de saída onde os *pixels* referentes aos pontos do objeto recebem o valor 1 (branco) e os *pixels* referentes ao fundo recebem o valor 0 (preto).

Histograma de uma Imagem Digital

A distribuição pela frequência de ocorrência dos níveis de cinza de uma imagem digital é denominada de histograma desta imagem. Assim, uma imagem possui uma faixa de níveis de cinza no intervalo $[0, ..., L]$, seu histograma $h(r_k) = n_k$ é uma função discreta, em que n_k é o número total de ocorrência dos níveis de cinza r_k dentro daquele intervalo.

Uma operação usual realizada sobre um histograma, com o objetivo de realce, é a limiarização. Esta consiste num mapeamento de *pixels* de uma imagem original $f(x, y)$ em outra imagem $g(x, y)$, por exemplo, dado pela equação

$$g(x,y) = \begin{cases} a, \ \textbf{se } f(x,y) \leq L, \\ b, \ \textbf{caso contrário,} \end{cases} \tag{2.1}$$

onde a, b e L são valores fixos, sendo que L é denominado de limiar, ou seja, é um valor limite que altera as propriedades dos *pixels* da imagem resultante $g(x, y)$, a partir da imagem original.

2.3.2 Segmentação por Bordas

A detecção de bordas é realizada pela descontinuidade na imagem, ou seja, de *pixels* em que o nível de cinza varia de repente em relação aos seus vizinhos. As duas técnicas mais comuns para detectar pontos de contorno são a localização de pontos de alto gradiente (usando a primeira derivada), ou cruzamentos de zero do Laplaciano(segunda derivada).

Detecção de Bordas Usando a Primeira Derivada

Uma vez aplicada à imagem um operador de derivada como o Sobel, Roberts, etc realiza-se uma classificação dos *pixels* da imagem segundo se são ou não pertencentes aos contornos.

Operador de Sobel

O operador de Sobel é uma técnica baseada em gradiente para a detecção de bordas de tal forma que as variações de intensidade prevalecem sobre as áreas de intensidade constante.

O trabalho executado por Sobel pode ser ilustrado simplesmente por funções unidimensionais, tais como o mostrado na Figura 2.23. A função $f(x)$ pode representar a variação da intensidade que ocorre em uma borda não tao abrupta. Se a primeira derivada é aplicada a esta função (que é o mesmo que a aplicação do gradiente, porque existe apenas na direção x), o resultado é outra função $f'(x)$, cujos valores máximos são fornecidos na região em que produz um aumento na intensidade. A primeira derivada e o operador de Sobel, por extensão possuem como propriedade destacar as áreas que não são constantes, e isso se traduz na capacidade de detecção de bordas nas funções bidimensionais, tais como imagens digitais.

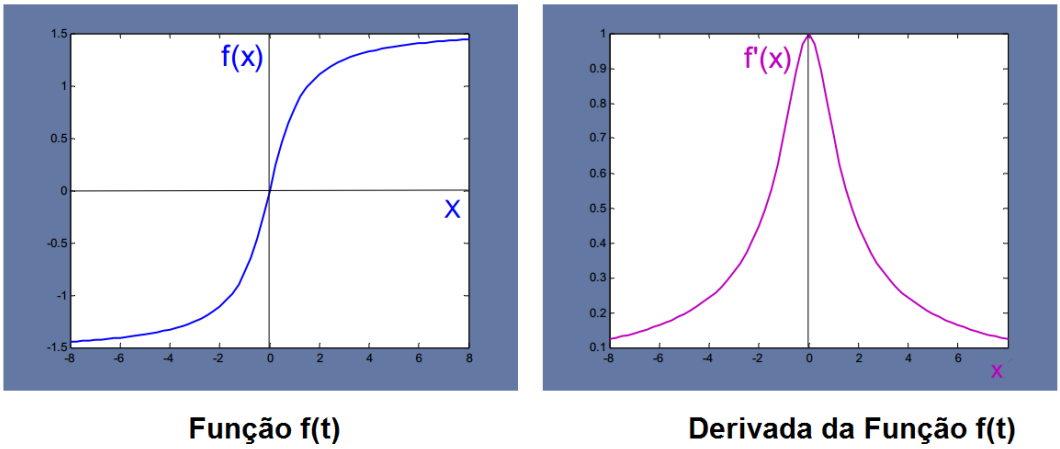


Figura 2.23: Função e seu Derivada.

Embora a derivada está associado à continuidade e as imagens digitais são discretos, pode ser aproximada assumindo que a imagem é apenas uma função continua amostrada em *pixels* que a definem.

Com base nesta abordagem, o operador Sobel permite obter para cada ponto (*pixel*) um vetor que define a direção em que ocorre o aumento da intensidade máxima possível

ou maior gradiente (preto para branco) entre este ponto e aqueles ao seu redor onde a magnitude aumento. Isto implica que o resultado da aplicação do operador Sobel sobre uma região de intensidade constante é um vetor 0, enquanto a aplicação sobre a borda produz um vetor perpendicular à direção deste cujo sentido vai do mais escuro para o mais claro. As máscaras que executam esta operação de detecção de bordas, de acordo com a direção na qual o gradiente é aplicado, são as seguintes G_x é aplicada na direção X, que corresponde às colunas da imagem, enquanto que G_y é aplicada em Y, direção na qual avança as filas.

$$G_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad G_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} \tag{2.2}$$

2.4 Operações pós-processamento

Muitas vezes o resultado da segmentação não é adequado, sendo necessária para se corrigir as imagens binárias resultantes da segmentação a aplicação de procedimentos de pós-processamento, como a separação de objetos mais complexos. Tais procedimentos são geralmente implementados através de operações lógicas e operações morfológicas.

2.4.1 Operações lógicas

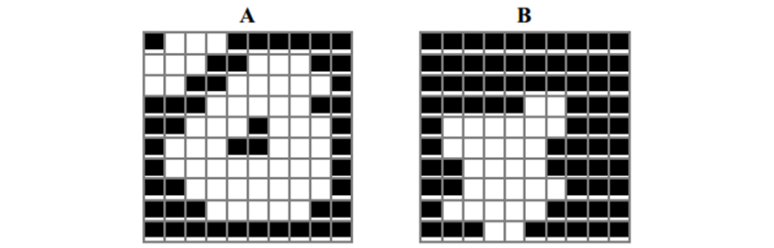
As operações lógicas são operações pontuais entre imagens binárias, realizadas por operadores lógico que varrem as imagens de entrada, operando *pixel* a *pixel*, gerando uma imagem de saída onde cada *pixel* é preservado ou invertido. As três operações lógicas básicas são o complemento (NOT), a interseção *(AND)* e a união *(OR)*, a partir das quais podem ser definidas qualquer outra operação lógica. A figura 2.24 mostra, a a partir de duas imagens de entrada (A e B), as três operações lógicas básicas com suas respectivas tabela de verdade.

A operação de complemento *(NOT)*, também chamado "não", inverte todos os *pixels* da imagem de entrada, gerando uma imagem de saída que é o seu negativo. A operação de interseção *(AND)*, também chamada "e", faz a interseção das duas imagens de entrada, produzem uma imagem de saída onde são brancos somente os *pixels* que são brancos em ambas imagens de entrada. A operação de união *(OR)*, também chamada "ou", realiza a união de duas imagens de entrada, produzindo uma imagem de saída onde são brancos somente os *pixels* que são brancos em pelo menos uma das imagens de entrada.

Outra operação logica, derivada dessas operações básicas, também bastante utilizada, é o "ou exclusivo"*(XOR)*. A operação de "ou exclusivo"*(XOR)* gera uma imagem de saída onde são brancos somente os *pixels* que são brancos em somente uma das imagens de entrada. De acordo com a notação estabelecida, tem-se que:

$$\{A \text{ XOR } B\} = \begin{cases} [A \textbf{ OR } B] \\ AND \\ [\textbf{NOT}(A \textbf{ AND } B)] \end{cases} \tag{2.3}$$

A Figura 2.25 mostra a operação XOR e a tabela de verdade que a define, a partir das mesmas duas imagens de entrada (A e B) da Figura 2.24



A	NOT A
0	1
1	0

A	B	A AND B
0	0	0
0	1	0
1	0	0
1	1	1

A	B	A OR B
0	0	0
0	1	1
1	0	1
1	1	1

Figura 2.24: Operações lógicas básicas *(NOT, AND e OR)*.

2.4.2 Operações morfológicas

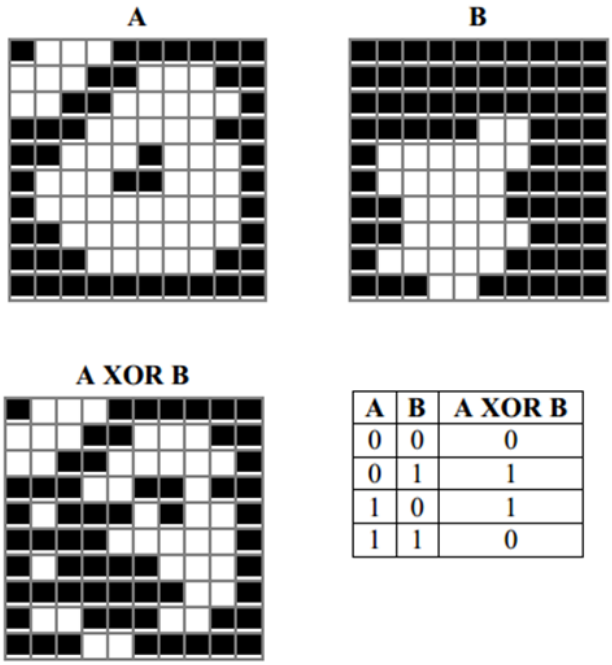
Assim como as operações lógicas são derivadas de apenas três operações lógicas básicas (NOT, AND e OR), as operações morfológicas também tem sua base, sendo geralmente definidas a partir de duas operações morfológicas de propagação, a erosão e a dilatação.

A erosão e dilatação são operações orientadas pela vizinhança. A vizinhança, neste caso, é definida por uma pequena imagem binária, denominada elemento estruturante, que varre a imagem de entrada, preservando ou invertendo o *pixel* central da vizinhança, na imagem de saída, em função de seus vizinhos.

A erosão e dilatação são operações de propagação opostas. como seus nomes sugerem, elas respectivamente, fazem os objetos encolher ou crescer. Na dilatação para cada *pixel* preto na imagem de entrada, se houver pelo menos um vizinho branco, o *pixel* é invertido na imagem de saída. Assim, os objetos crescem em área, podendo ate se unir, e o fundo e os buracos nos objetos diminuem, sendo até mesmo eliminados. Na erosão, para cada pixel branco na imagem de entrada, se houver pelo menos um vizinho branco, o *pixel* é invertido na imagem de saída. De modo que os objetos diminuem ou são eliminados e o fundo e os buracos crescem.

A Figura 2.26 mostra a erosão e dilatação de uma imagem (A), de 10×10 *pixels* por um elemento estruturante (E), de 3×3 *pixels*.

O elemento estruturante (E) da Figura 2.26 é chamado de conectividade 4 (Figura 2.27-a). Ele define a vizinhança como 3×3, sendo considerados vizinhos de *pixel* central somente os 4 *pixels* adjacentes a ele lateral e verticalmente. um outro elemento estruturante, também muito utilizado, é o elemento de conectividade 8 (Figura 2.27-b), que define a vizinhança como 3×3, mas considera todos os 8 *pixels* adjacentes como vizinhos do *pixel* central. contudo, podem ser definidos elementos estruturantes das mais variadas



A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0

Figura 2.25: Operação lógica *XOR*.

formas e tamanhos, sendo a escolha do mais adequado somente em função do problema. Nos exemplos que seguem, como na Figura 2.26, é utilizado o elemento de conectividade 4.

A partir das operações lógicas básicas (NOT, AND e OR) e das operações morfológicas de erosão e dilatação, são definidas outras operações morfológicas simples porém importantes, como o preenchimento de buracos, a eliminação de objetos que tocam na borda da imagem e a extração de objetos marcados.

Dada uma imagem (A) e um elemento estruturante (E), uma imagem contendo somente as bordas dos objetos de A pode ser obtida por uma operação morfológica de extração de bordas (Figura 2.28), definida como:

$$b\ (\mathrm{A})\ = \{[NOT(A \ominus E)]\ AND\ \boldsymbol{A}\ \} \tag{2.4}$$

A partir desta operação de extração de bordas, pode-se definir a operação morfológica de preenchimento de buracos, mostrada na Figura 2.29, definida como:

$$X_0 = b(Ib) \tag{2.5}$$

$$X_k = \{[X_{k-1}\ \mathbf{AND}\ (\ \mathbf{NOT}\ A)] \oplus E\} \tag{2.6}$$

$$k \in X^* \tag{2.7}$$

$$fh(A) =\ \mathbf{NOT}\ X_k | (X_k = X_{k-1}) \tag{2.8}$$

onde *Ib* é uma imagem com todos os *pixels* brancos, da mesma dimensão que A.

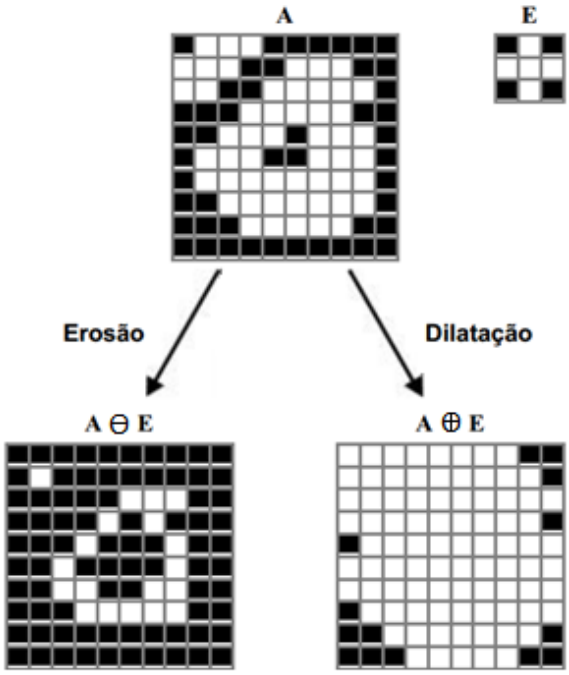


Figura 2.26: Erosão e dilatação.

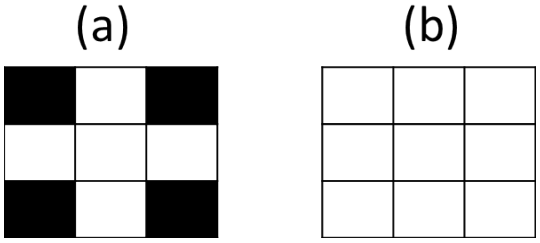


Figura 2.27: Elemento estruturante 3×3: (a)conectividade 4; (b) conectividade 8.

2.5 Revisão Bibliográfica

Em anos recentes, mais metodologias são propostas em relação à conversão de vídeo monoscópica para o vídeo estereoscópico. A criação de conteúdo estéreo exige o uso de técnicas de síntese de vista para obter as duas vistas necessárias. Nesta seção vamos descrever rapidamente os principais métodos de conversão incluindo um método mais específico para conteúdo esportivo.

Têm-se diferentes métodos de conversão 2D para 3D. Alguns métodos reconstroem mapas de textura onde o usuário especifica as posições de superfícies (como regiões planas e de curvatura) [26] ou posições de objetos que são traçados manualmente em quadros chaves de um vídeo [27]. Outros métodos são automáticos e utilizam conhecimento a priori da cena como a localização e orientação. Estes métodos são: *Make3D* [28] e *Automatic photo pop-up* [29] os quais funcionam por construção de modelos estáticos de cenas 3D, segmentação de imagens e aplicação de relações entre planos. Estes abordagens são limitadas e não oferecem estabilidade temporal.

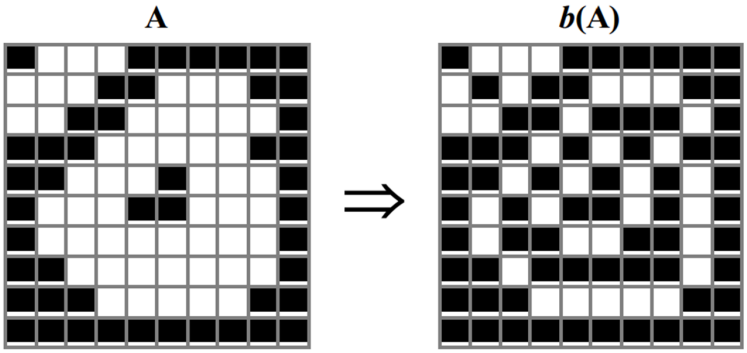


Figura 2.28: Extraçao de bordas.

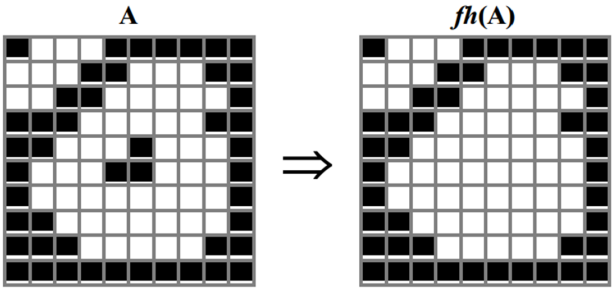


Figura 2.29: Preenchimento de buracos *(hole filling)*.

Uma técnica muito usada é a DIBR (do ingles*Depth Image Based Rendering*). O DIBR precisa de uma sinal de profundida que é o chamado mapa de profundidade , além do sinal de textura [30].

No entanto, para a conversão de vídeo, a estabilidade temporal de mapas de profundidade é muito importante para alcançar a estabilidade temporal construí um panorama de fundo [2]

2.5.1 Modelagem de visualização única de cenas em forma livre.

Neste método [26] se apresenta uma nova abordagem para reconstruir de forma livre mapas de textura e modelado de uma cena 3D a partir de uma única pintura ou imagem (Figura 2.30). O usuário especifica posições de superfícies como "regiões planas", "regiões dobradas", "regiões descontinuas" e "regiões de curvatura" gerando uma superfície 3D que satisfaze as restrições. Este problema é formulado como um problema de otimização restrita variacional [31]. Ao especificar cada restrição, o sistema recalcula e exibe a reconstrução em tempo real. Uma característica fundamental de este método é que utiliza uma técnica de transformação hierárquica moderna que acelera a convergência num plano não uniforme. A técnica é interativa e atualiza o modelo em tempo real como as restrições são adicionadas permite a reconstrução rápida de modelos de cenas realistas.

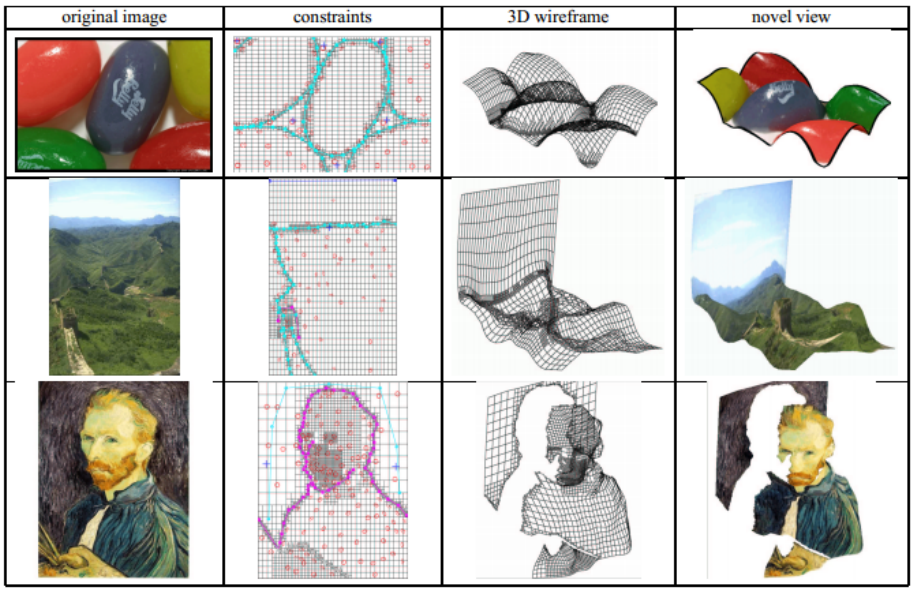


Figura 2.30: Exemplos de modelagem única em diferentes cenas. Da esquerda para a direita, as colunas mostram: as imagens originais, restrições especificados pelo usuário em redes adaptativas, estrutura de renderização 3D e renderização de textura [26].

2.5.2 Video Trace

Este trabalho [27] apresenta um sistema interativo que gera modelos realistas 3D de objetos de um vídeo que podem ser inseridos em um jogo de vídeo, um ambiente de simulação ou de outra sequência de vídeo (Figura 2.31). O usuário interage com *Video Trace* traçando a forma do objeto a ser modelado sobre um ou mais quadros de vídeo. Ao interpretar o esboço desenhado pelo usuário em função das informações 3D obtido a partir de técnicas de visão por computador, um pequeno número de interações simples de 2D pode ser usado para gerar um modelo 3D realista. Cada uma das operações de esboço fornece uma maneira intuitiva e poderosa de modelagem do vídeo, e executa com suficiente rapidez para ser usado interativamente. A retroalimentação imediata permite que o usuário possa modelar rapidamente as partes da cena que são de inteires e com nível de detalhe requerido.

2.5.3 Make3D

Neste método [28], tanto a localização como a orientação 3D de pequenas regiões planas na imagem são deduzidos usando um Campo Aleatório de Markov (MRF) [32]. No método a relação entre as características da imagem e a localização/orientação dos planos esta baseado na aprendizagem. Também são achadas as relações entre varias partes da imagem usando aprendizagem supervisada [28]. O indicio básico aqui é chamada "superpixel", que a sua vez está relacionada à localização/orientação dos planos. As saídas desejadas utilizadas durante o treinamento são obtidas por um escâner laser 3D. Um banco de dados que consiste em vários tipos de cena é criado inicialmente, e então o mapa de profundidade destas cenas é gerado pelo escâner 3D. Dependendo da probabilidade de

Figura 2.31: A imagem mostra: um quadro de vídeo sequência, o traçado parcial do modelo, o modelo final sobreposto ao vídeo, e o resultado de renderização do modelo final de volta na sequência original [27].

mapa de superpixels e o mapa de profundidade de escâner, é obtida a reconstrução da estrutura de uma imagem 2D (Figura 2.32).

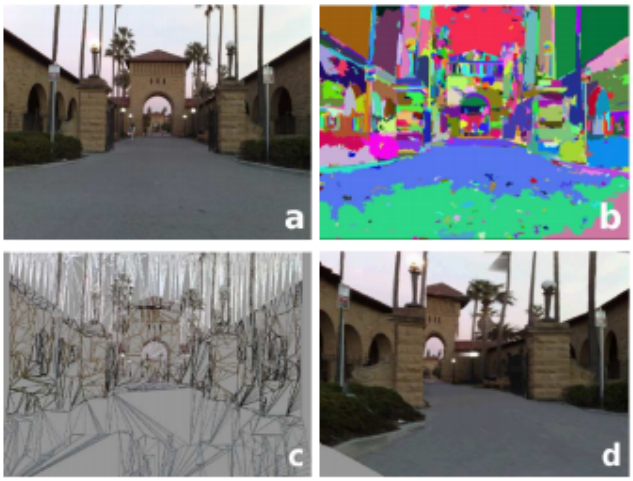


Figura 2.32: (a)Imagem original. (b) Segmentação da imagem para obter "Superpixels". (c) O modelo 3D previstas pelo algoritmo. (d) Imagem do modelo 3D texturizado [28].

2.5.4 Pop-up automático de imagens

Este trabalho [29] apresenta um método totalmente automático para criar um modelo 3D a partir de uma única imagem. O modelo é composto por vários mapas de texturas e tem a complexidade de ilustração para crianças, como o típico livro “pop-up” (Figura

2.33). O objetivo principal é que em vez de tentar recuperar a geometria precisa, se modela estatisticamente classes geométricas definidas por suas orientações na cena. O método assume que as imagens de entrada são de cenas ao ar livre, naturais e artificias (edifícios) e rotula as regiões da imagem em categorias grosseiras: "chão", "céu" e "vertical" (áreas que são perpendiculares ao chão). Essas etiquetas são utilizadas para "cortar e dobrar" a imagem num modelo "pop-up" usando um conjunto simples de premissas. A cor, textura, localização da imagem e características geométricas são dicas uteis para determinar estes rótulos.

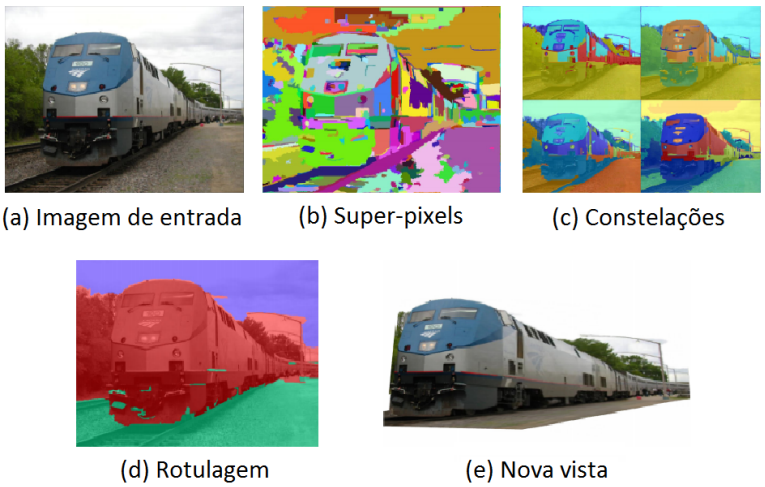


Figura 2.33: Para obter estatísticas úteis para o modelado das classes geométricas, primeiro temos que encontrar de maneira uniforme o as regiões marcadas por superpixels (b) e agrupando-os em constelações múltiplas (c). Então podemos gerar conjunto de estatísticas e rotular a imagem com base em modelos aprendidos a partir de imagens de treinamento. A partir destes rótulos, podemos construir um modelo simples 3D (e) da cena. Em (b) e (c), as cores distinguem-se entre as regiões separadas; em (d) As cores indicam as etiquetas geométricos: chão, verticais e céu [29].

2.5.5 Estereoscopia psicológica e fisiológica

Normalmente, os seres humanos percebem profundidade com dois olhos. Com um único olho, é difícil ou até mesmo impossível perceber informação de profundidade [33]. Assim neste método [30] tenta-se criar uma nova imagem a partir do original. Com as duas imagens, as pessoas podem construir a estrutura 3D facilmente baseado na psicologia do humano e função fisiológica. Como se sabe, a imagem estéreo é criada no cérebro através de dois olhos. Há dois tipos de estereoscopia em seres humano. A estereoscopia psicológica, o qual está baseado na memoria visual e experiência. Significa que quando as pessoas veem uma imagem eles podem entender o conteúdo desta imagem com os indícios existentes na imagem e não é difícil que as pessoas saberem a relação entre os objetos diferentes. Estes indícios são algum tipo de experiência e memória em cérebro humano, que é reunido em um período longo. Assim os seres humanos podem perceber a profundidade e posição de objetos numa imagem com a própria experiência e memoria deles. O outro tipo é a estereoscopia fisiológica; este está baseado na função física e estrutura dos olhos

do humano. Disparidade binocular [34] é o elemento mais importante de estereoscopia fisiológica. Foi provado que depois de excluir todos os elementos psicológicos, é possível que um jogo de excitações visuais possa criar a sensação de profundidade que sente com dois olhos sob a condição de disparidade binocular. Como a disparidade binocular tem o efeito mais forte na estereoscopia, é o elemento mais importante a ser considerado na criação de imagens estéreo. Este método tenta fazer que todo *pixel* da imagem troque só na direção horizontal. Isto introduzirá as disparidades binoculares em uma única imagem e conduzira a algum sentido estereoscópico. Este tipo de paralaxe binocular também pode fortalecer a estereoscopia psicológico.

2.5.6 Conversão de 2D para 3D de conteúdo esportivo usando panoramas

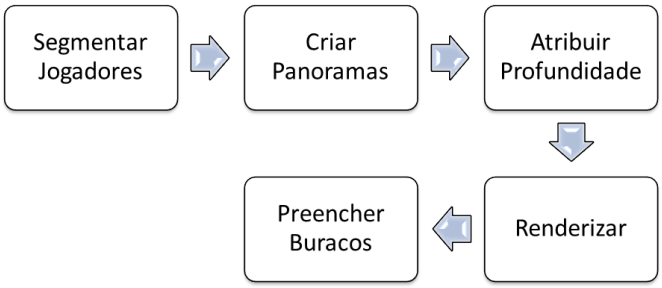


Figura 2.34: Visão geral do método de processamento para a conversão de vídeo 2D para 3D [2].

Este trabalho [2] apresenta um sistema para criar automaticamente vídeo estereoscópico de alta qualidade a partir de um vídeo monoscópico de conteúdo esportivo tendo como vantagem conhecimento prévio de contexto, tais como a geometria e aparência do campo esportivo, tamanho dos jogadores e orientação do plano de fundo. A ideia básica de este método (Figura 2.34) é separar partes estáticas e dinâmicas da cena, onde a parte estática é o fundo e a parte dinâmica são os jogadores. Depois de fazer a segmentação se constrói um panorama de fundo para cada cena (uma serie de *frames* sequenciais pertencentes à mesma cena) utilizando uma abordagem de mosaico clássica com uma câmara de rotação fixa para as homografias. A partir disso é criado um mapa de profundidade para todo o panorama usando as suposições sobre a estrutura do campo esportivo. Uma vez obtido o mapa de profundidade do panorama se gera os mapas de profundidade do fundo para cada *frame*, eles podem ser calculados por uma projeção inversa usado na homografia anterior, em seguida os valores de profundidade para os jogadores são atribuídos em cada frame do plano segmentado como painéis, assumindo que a câmera esta alinhada verticalmente e que os jogadores estão em estreito contato com o solo, a profundidade para os jogadores é atribuída a partir de ponto mais baixo da região segmentada no plano de profundidade. Ambiguidades na segmentação são corrigidas para não causar erros perceptíveis, dando-nos um mapa de profundidade final de cada frame de entrada, e produzindo cenas 3D sintetizadas (Figura 2.35).

Figura 2.35: Anaglyph imagem de uma cena estéreo resultante [2].

Capítulo 3

Conversão de 2D a 3D

3.1 Descrição do Método

Nesta seção, propomos um novo método para converter vídeos monoscópicos de futebol para vídeos estereoscópicos, baseado em uma proposta anterior [2]. A maioria de métodos de conversão de 2D a 3D utilizam como informação o movimento. No entanto, no caso de vídeos de futebol, as cenas mudam muito rápido e há muitos objetos pequenos (por exemplo, jogadores, bola, etc). Assim, elaboramos um novo método para converter o vídeo de futebol em vídeo estereoscópico utilizando suas próprias características como a geometria do "campo" de jogo, tamanho do jogador e o conhecimento do fundo. A Figura 3.1 ilustra o método proposto onde a idéia básica é separar a parte estática da parte dinâmica da cena e processar cada uma usando algoritmos específicos [35]. Em seguida cria-se um mapa de profundidade, o qual será utilizado para gerar um par estéreo a partir de uma entrada monoscópica. O método funciona para sequências de imagens panorâmicas e transmissões de vídeo com câmeras que não são fixas e se movem para acompanhar os jogadores.

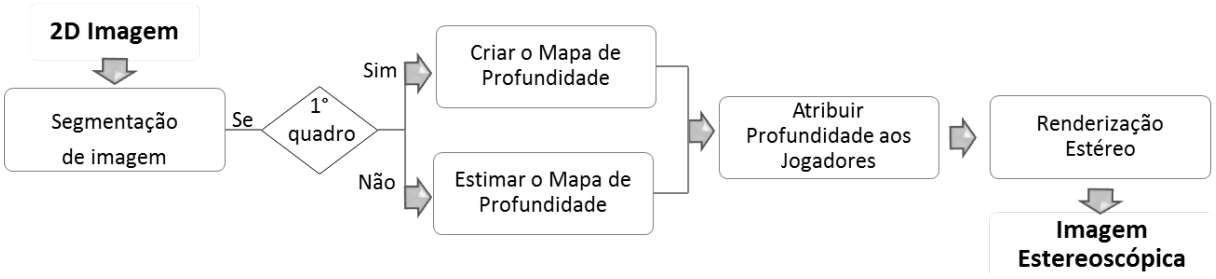


Figura 3.1: Procedimento geral do método proposto para a conversão de 2D para 3D.

Como indicado na Figura 3.1, em primeiro lugar, apresentamos a segmentação da imagem, onde a parte estática é o fundo ou "campo" e a parte dinâmica são os jogadores. No seguinte passo, para o primeiro quadro da sequência de vídeo é criada um mapa de profundidade. Para os próximos quadros o mapa de profundidade é estimado a partir de mapas de profundidade anteriores, com isso, evita-se a cintilação (*flickering*). No passo seguinte, utilizamos o resultado da segmentação dos jogadores para atribuir profundidade a cada um deles. Finalmente, apresentamos a renderização estéreo.

As seguintes condições não foram contempladas:

- Vídeos produzidos a partir de várias câmeras que alternam dinamicamente as cenas.
- Vídeos que mostram mudanças rápidas do fundo.
- Vídeos onde um das equipes tenha a camisa verde (cor semelhante à grama)
- Vídeos que tem um "campo" com a grama ocupando uma porcentagem menor de 60% da imagem.
- Vídeos com *zoom* óptico [1].

3.2 Segmentação dos Jogadores

Os jogadores e a bola são os objetos mais importantes em vídeos de futebol. A segmentação deles é motivada por várias aplicações, tais como detecção de eventos, análise tática, sumarização automática e compressão baseada em objetos [36].

O "campo" de jogo nos fornece a informação mais relevante e confiável para a análise de cenas de futebol. Uma vez que as câmeras costumam se concentrar nos jogadores que conduzem a bola, a maioria dos quadros de vídeo capturam apenas uma parte do "campo". Portanto, na primeira etapa da análise de jogo de futebol, extraímos a região do "campo" de futebol em si (fundo). Em uma segunda análise, combinamos várias ferramentas para extrair automaticamente os jogadores e as linhas de marcação do "campo".

3.2.1 Detecção dos *Pixels* do "campo"

A partida de futebol é normalmente acontece num "campo" de grama. Neste trabalho supõe-se que o "campo" tem uma cor verde uniforme e ocupa um área grande na imagem (uma área maior ou igual a 60% da imagem). A fim de detectar os jogadores e a bola, um primeiro passo útil é detectar os *pixels* que formam o "campo". Assim como foi proposto em [35], empregamos uma técnica de aprendizagem simples, baseado em histogramas para detectar os *pixels* do "campo". Esta estrutura foi proposta inicialmente por Jones [37] para a detecção de pele.

A detecção dos *pixels* do "campo" (fundo) e *pixels* de "não-campo" (jogadores) são aprendidas da seguente forma:

- A partir do primeiro quadro do vídeo são gerados manualmente duas imagens, uma com *pixels* marcados como "campo" e a outra com *pixels* marcados como "não-campo", como se mostra na Figura 3.2.

[1] O *zoom* óptico é um tipo de *zoom* que consegue aumentar ou diminuir o enfoque da imagem utilizando lentes da câmera fotográfica ou vídeo, alterando a distância focal.

Figura 3.2: *Pixels* marcados como "campo" e "não-campo": (a) imagem original; (b) *pixels* marcados manualmente como "campo"; (c) *pixels* marcados manualmente como "não-campo".

- Paras os *pixels* marcados como "campo" e "não-campo" realiza-se histogramas nos três componentes de cor (R, G e B), cada uma das três componentes esta dividida em um número de "cestos" ($Nbins$) [2]. Cada "cesto" armazena um número inteiro de contagens de número de vezes que o valor da cor ocorre na imagem. Um exemplo destes histogramas é ilustrado na Figura 3.3 que esta dividido em 128 "cestos".

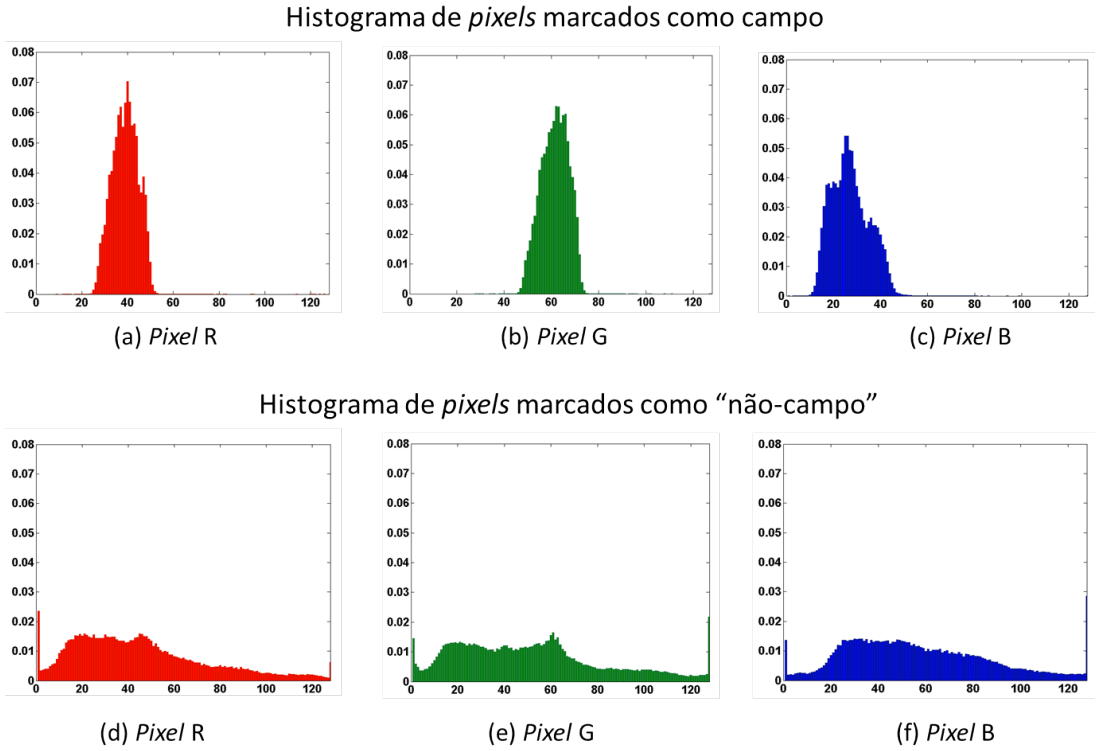


Figura 3.3: Histograma para a detecção dos *pixels* do "campo". (a),(b) e (c) são os histogramas de *pixels* marcados como "campo" em R, G e B respectivamente. (c),(d) e (f) são os histogramas de *pixels* marcados como "não-campo" em R, G e B respectivamente.

[2] Tipicamente o número de "cestos" utilizados para os histogramas são 32, 64, 128 e 256.

- Determina-se um vetor de probabilidades para cada histograma, onde cada probabilidade é o resultado da relação do número de *pixels* contidos em um "cesto" pelo número total dos *pixels* dos histogramas de "campo" e "não-campo" respectivamente. Os vetores de probabilidades para os histogramas são determinados pelas seguintes equações:

$$\begin{aligned}
P(R_j \in campo) &= \frac{c(R_j)}{T_c} \qquad & P(G_j \in não\text{-}campo) &= \frac{n(R_j)}{T_n} \\
P(G_j \in campo) &= \frac{c(G_j)}{T_c} \qquad & P(R_j \in não\text{-}campo) &= \frac{n(G_j)}{T_n} \\
P(B_j \in campo) &= \frac{c(B_j)}{T_c} \qquad & P(B_j \in não\text{-}campo) &= \frac{n(B_j)}{T_n}
\end{aligned} \tag{3.1}$$

  onde $j = 1, ..., Nbins$. $P(R_j \in campo)$, $P(G_j \in campo)$ e $P(B_j \in campo)$ são as funções de distribuição de probabilidade de ocorrência de cada nível (j-ésimo "cesto") do espaço de cor R, G e B respectivamente, dos histogramas de *pixels* marcados como"campo". $c(R_j)$, $c(G_j)$ e $c(B_j)$ são as funções que determinam os números totais de *pixels* contidos no j-ésimo "cesto" do espaço de cor R, G e B respectivamente, dos histogramas de *pixels* marcados como"campo". De forma similar, $P(R_j \in não\text{-}campo)$, $P(G_j \in não\text{-}campo)$ e $P(B_j \in não\text{-}campo)$ são as funções de distribuição de probabilidade de ocorrência de cada nível (j-ésimo "cesto") do espaço de cor R, G e B respectivamente, dos histogramas de *pixels* marcados como "não-campo". $n(R_j)$, $n(G_j)$ e $n(B_j)$ são os números totais de *pixels* contidos no j-ésimo "cesto" do espaço de cor R, G e B, respectivamente, dos histogramas de *pixels* marcados como "não-campo". T_c e T_n são as contagens totais de *pixels* contidas nos histogramas de *pixels* marcados como "campo" e "não-campo" respectivamente.

Depois de ter feito o aprendizagem dos *pixels* do "campo" obtém-se um classificador de *pixels* através do método padrão de relação de probabilidade. Então, um valor de cor do espaço RGB em particular é rotulado como "campo" se as três equações seguintes são verdadeiras.

$$\frac{P(R_j \in campo)}{P(R_j \in não\text{-}campo)} \geq \theta \tag{3.2}$$

$$\frac{P(G_j \in campo)}{P(G_j \in não\text{-}campo)} \geq \theta \tag{3.3}$$

$$\frac{P(B_j \in campo))}{P(B_j \in não\text{-}campo)} \geq \theta \tag{3.4}$$

onde $\theta \geq 0$ é um limiar que pode ser ajustado para obter mais detecções de falsos positivos que falsos negativos. Após o processo da classificação dos *pixels* do "campo" é obtido a máscara binária, onde o valor 0 representa os *pixels* do "campo" e o valor 1 os *pixels* de "não-campo".

A Figura 3.4 mostra o resultado do método de detecção de "campo". A imagem original é apresentada à esquerda. Os *pixels* detectados como "campo" são mostrados no lado direito (máscara binária). A operação de abertura morfológica será aplicada após a detecção de *pixels* de "campo", a fim de remover pequenos falsos positivos. Pode-se observar na Figura 3.4-(b), que o método tem sucesso na detecção de *pixels* do "campo" sobre variações significativas na cor (faixas escuras e claras) e iluminação (áreas iluminadas e sombreadas).

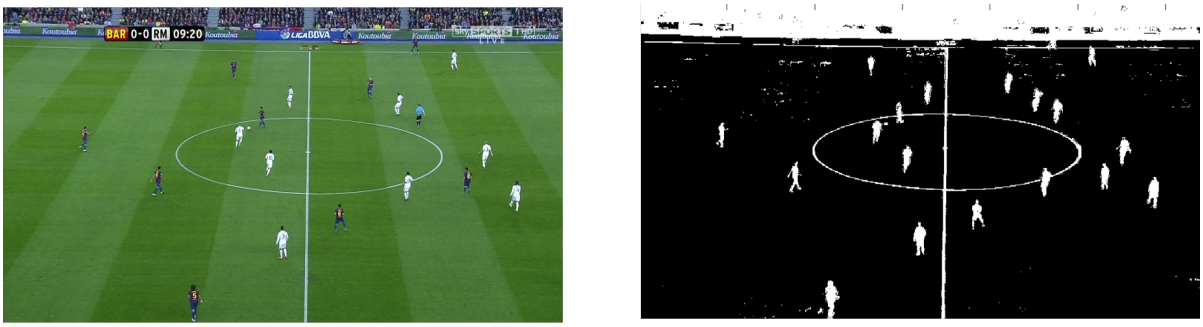


Figura 3.4: Resultado da detecção do "campo" de jogo: (a) Quadro original e (b) *pixels* detectados do "campo".

3.2.2 Detecção dos *pixels* de "não-campo"

Dado que os *pixels* classificados como "campo" já foram achados o restante, que são os *pixels* "não-campo", podem ser ainda sub-classificados como:

- Jogadores
- Bola
- Público
- Linhas do campo

Após ter passado pelo detector de *pixels* do "campo" e ter obtido uma máscara binária como ilustrado na Figura 3.4-(b) realiza-se a extração de "não-campo" como ilustra a Figura 3.5 para o qual é feita a filtragem morfológica (erosão, e em seguida a dilatação) aplicada para eliminar ruído e os falsos positivos.

Extração do público

Para a detecção do público são utilizados operações morfológicas e lógicas conforme os passos seguintes:

1. Realizar a operação morfológica de preenchimento de buracos ou regiões na máscara binária, representada pela Figura 3.6-(a). O objetivo desse passo é eliminar o ruido existente na área do público. Como resultado obtém-se a Figura 3.6-(b).

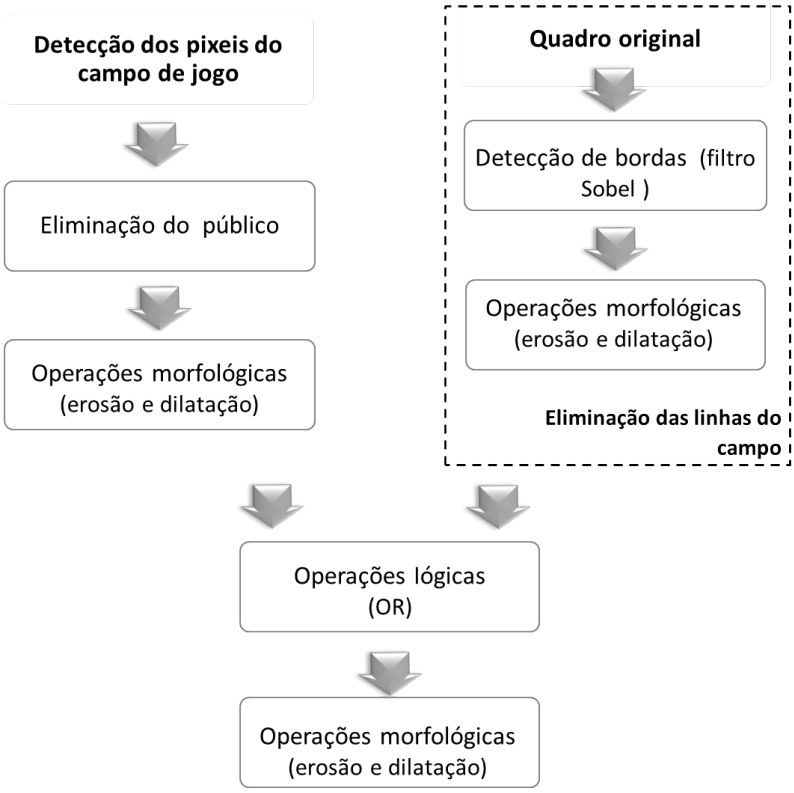


Figura 3.5: Extração de ruídos e falsos positivos na máscara binária.

2. Seguidamente inverter os *pixels* do resultado anterior com o objetivo que os jogadores e as linhas do campo representem buracos e dessa forma poder aproveitar o algoritmo de preenchimento de buracos do passo anterior (passo 1). O resultado desse passo é ilustrado na 3.6-(c).
3. Realizar o processo morfológico de preenchimento de buracos no resultado anterior (passo 2). Como resultado obtém-se a Figura 3.6-(d).
4. Finalmente, realizar a operação lógica de "and" entre o resultado do passo 3 e passo 2 com o objetivo de obter uma imagem com jogadores e linhas do campo como ilustrado na Figura 3.6-(e).

O processo anterior pode ser representado pela seguinte equação:

$$P = fh\left(NOT\left(fh\left(Mb\right)\right)\right) \;\; AND \;\; fh\left(Mb\right) \tag{3.5}$$

onde Mb é a máscara binária, fh é a operação morfológico de preenchimento de buracos explicado na Subseção 2.4.2, NOT é a operação lógica que inverte os valores dos *pixels*, XOR é a operação logica de "ou exclusivo" e P é o resultado.

Extração de linhas do campo

Com o objetivo de detectar os contornos das linhas no campo de jogo, foi utilizado o filtro Sobel [24] na imagem original (em escala de cinza). Como explicado anteriormente na

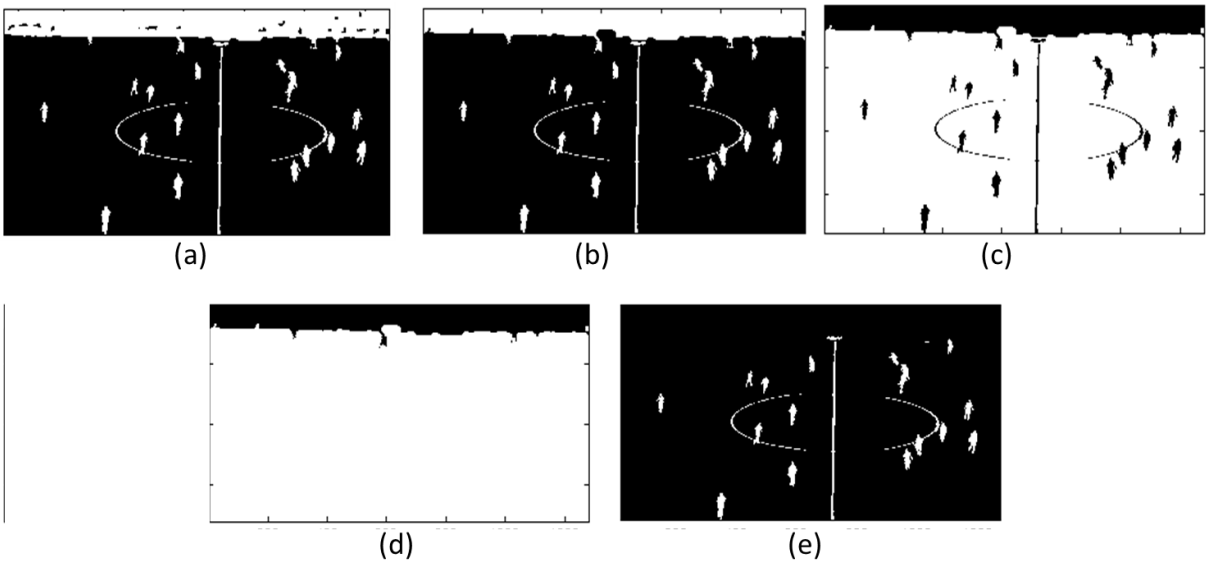


Figura 3.6: Processo de extração do público: (a) máscara binária; (b)operação do preenchimento de buracos para limpar ruídos no publico; (c) inversão dos *pixels*; (d) operação do preenchimento de buracos limpar jogadores e linhas do campo; (e) máscara binária sem público.

Subsecção 2.3.2, este algoritmo permite detectar numa imagem os contornos dos objetos como ilustrado na Figura 3.7.

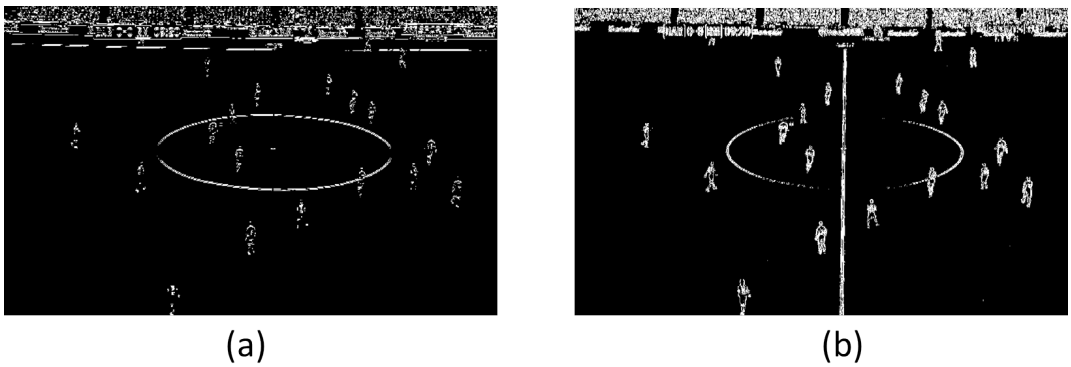


Figura 3.7: Aplicação de filtro Sobel: (a) Resultado da imagem da Figura 3.4-(a) após aplicação do filtro de Sobel para detectar bordas horizontais e (b) para detectar bordas Verticais.

Na Figura 3.7 pode-se observar o resultado da aplicação do filtro de Sobel para a detecção do contorno das linhas do campo, mas além dos contornos das linhas são também detectadas outros contornos. Para evitar que outros objetos sejam detectados e possam ser confundidos com as linhas, tornou-se necessário encontrar uma forma de exclusão deste ruido conforme às seguintes equações:

$$Lv = V \ \ AND \ \ (NOT\,(D\,(H))) \tag{3.6}$$

$$Lh = H \ \ AND \ \ (NOT\,(D\,(E\,(V)))) \tag{3.7}$$

$$Lc = P \ \ AND\,(Lh \ \ OR \ \ Lv) \tag{3.8}$$

Onde P é a imagem sem público (Figura 3.6-(f)), H imagem de bordas horizontais (Figura 3.7-(a)), V imagem de bordas verticais (Figura 3.7-(b)), E operação morfológica de erosão com um elemento estruturante de 3×3 de conectividade 4, D operação morfológica de dilatação com um elemento estruturante de 10×10 de conectividade 8, Lv são as linhas verticais obtidas da Equação 3.6 como é ilustrada na Figura 3.8-(a), Lh são as linhas horizontais obtidas da Equação 3.7 como é ilustrada na Figura 3.8-(b) e por último Lc é o resultado da Equação 3.8, ilustrado na Figura 3.8-(c).

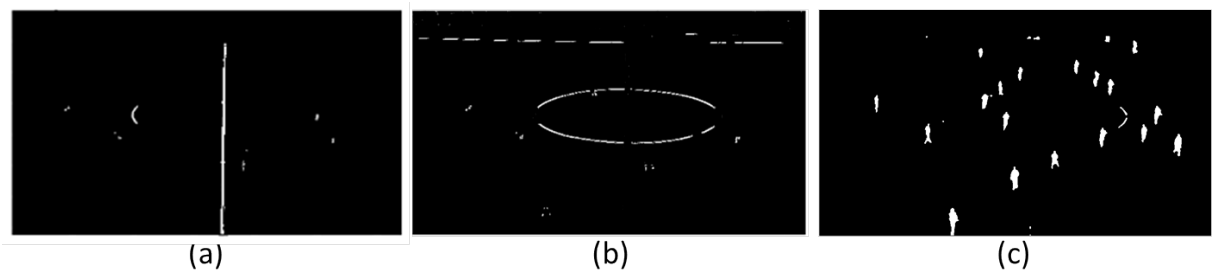


Figura 3.8: Processo de extração de linhas: (a) imagem de linhas verticais; (b) imagem de linhas horizontais e (c) imagem sem linhas de campo.

3.2.3 Análise de forma para limpeza de "não-campo"

Após eliminar o público e linhas do campo da máscara binária obtemos a imagem ilustrada na Figura 3.8-(c). Nesta, há várias regiões, cada uma tem propriedades como perímetro, área, excentricidade e etc. A partir destas informações, podemos distinguir o jogador e os segmentos de linhas que ainda restam do processo anterior.

As propriedades para cada região são dadas por alguns descritores de forma básica, tais como perímetro P, área A, comprimento do eixo maior C_L, comprimento de eixo menor C_S , arredondamento $F = P^2/(4\pi A)$ e excentricidade $E = C_L/C_S$.

Para que uma região seja considerada como jogador, tem que cumprir as seguintes condições:

- A área da região tem que variar entre os limites de 340 a 2000
- A excentricidade tem que variar entre os limites de 1.2 a 6.0
- O arredondamento tem que variar entre os limites de 1.4 a 9.0

Para que uma região seja considerada como bola, tem que cumprir as seguintes condições:

- A área da região tem que variar entre os limites de 60 a 150
- A excentricidade tem que variar entre os limites de 0.9 a 1.6
- O arredondamento tem que variar entre os limites de 1.0 a 1.4

Os limiares utilizados para que uma região seja considerada como jogador e bola foram determinadas empiricamente utilizando 100 quadros de 1280 *pixels* de diferentes vídeos, tendo em conta as limitações listados na Seção 3.1. A Figura 3.9 mostra o resultado da limpeza das regiões que não são considerados jogadores e bola, considerando suas propriedades geométricas. Este é o resultado final do processo de segmentação.

Figura 3.9: Resultado da limpeza de "não-campo" (resultado final da segmentação, jogadores e bola).

3.3 Modelo do Mapa de Profundidade

Neste trabalho propomos um método simples para a estimação do mapa de profundidade que possa ser utilizado em tempo real. As imagens com as quais trabalhamos são classificadas como "vista panorâmica sem ponto de desaparecimento" [22]. A geração de mapas de profundidade para imagens de "vista panorâmica sem ponto de desaparecimento" é realizada através da exploração de gradiente e mecanismos intuitivos da visão binocular (sinais perspectiva linear) como foi explicado na Subseção 2.1.4. Porém o mapa de profundidade pode ter mudança de gradiente no sentido vertical. Esta abordagem é projetado para a extração de profundidade qualitativa, que se espera que seja adequado para 3DTV.

O mapa de profundidade para a imagem (Figura 3.10-(a)) é definido por:

$$D(x,y) = x \left\lfloor \frac{Nniveis}{altura} \right\rfloor + Nivel_{inicial} \tag{3.9}$$

onde x varia no índice vertical (número de linhas na imagem), a variável *altura* representa a altura da imagem em *pixels*, $Nniveis$ representa o número de níveis de profundidade (para a imagem panorâmica toma o valor de 255) e $Nivel_{inicial}$ é o valor inicial do nível de profundidade (para a imagem panorâmica toma o valor de 0). A Figura 3.10-(a) mostra a imagem original classificada como "vista panorâmica sem ponto de desaparecimento" e a Figura 3.10-(b) mostra a geração de mapa de profundidade correspondente.

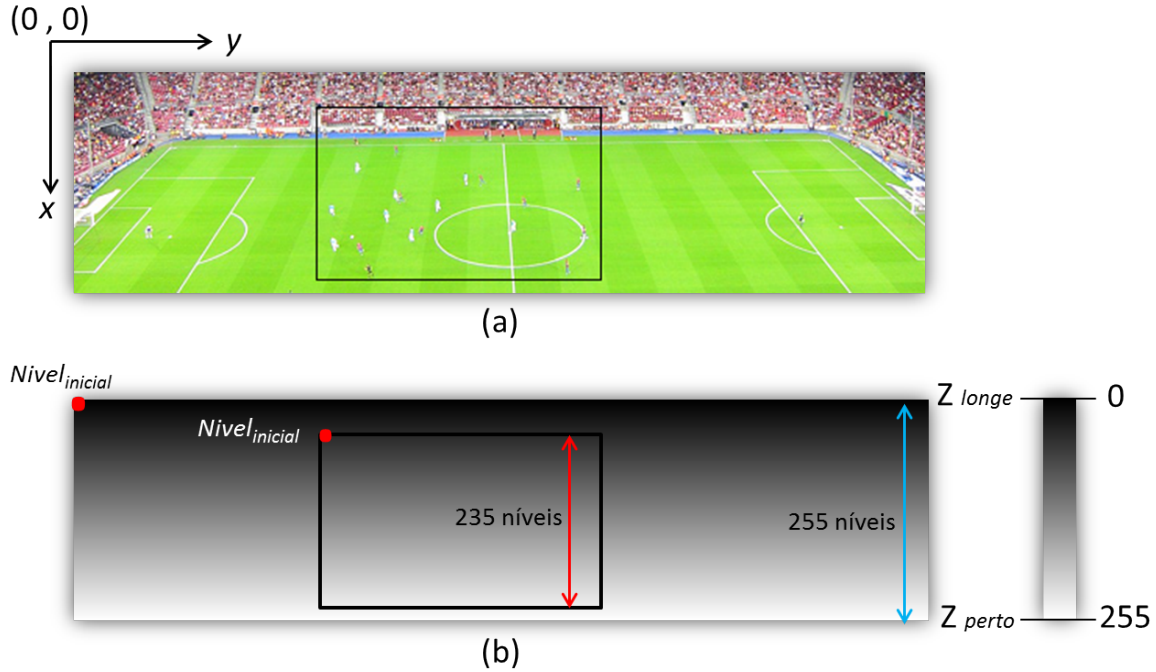


Figura 3.10: Geração do mapa de profundidade para "vista panorâmica sem ponto de desaparecimento". (a) imagem assumida como o intervalo de gravação do vídeo; (b) mapa de profundidade gerado.

Uma vez que as câmeras costumam se concentrar nos jogadores que conduzem a bola, a maioria dos quadros de vídeo capturam apenas uma parte do "campo" como se mostra na Figura 3.10-(a) marcado com o quadrado preto, estes quadros são obtidos pelo movimento da câmera da cima para abaixo e da esquerda para a direita e vice-versa. Logo, se o plano de fundo (imagem panorâmica assumida como intervalo de gravação de vídeo) é normalizada com valores de profundidade 0 e 255. Para um determinado instante de tempo os níveis de profundidade para o quadro (Figura 3.10-(b) indicado com o quadro preto) devem ser válidos, de tal forma, que dada a movimentação vertical da câmera os níveis não ultrapassem os limites de 0 e 255.

O mapa de profundidade para o primeiro quadro do vídeo (indicado com o quadrado preto na Figura 3.10-(a)) é obtido também com a Equação 3.9 ajustando as variáveis $Nivel_{inicial}$ e $Nniveis$. A variável $Nivel_{inicial}$ toma o valor de 10, este valor foi determinado empiricamente mediante a análise do movimento vertical de mais de 300 quadros, esse valor variará de acordo ao movimento vertical do quadro. O parâmetro $Nniveis$ toma o valor de 235, já que o quadro ocupa uma parte só da imagem como ilustrado na Figura 3.10-(b) marcado com o quadrado preto. Esse valor é obtido mediante a diferença de 255 níveis da imagem panorâmica com a variação do movimento vertical do quadro tanto acima como abaixo (20 níveis em promédio).

3.3.1 Propagação do Mapa de Profundidade

Um dos passos importantes na geração de vídeo 3D é de evitar a cintilação causado pela propagação de erros no mapa de profundidade, que é um efeito desagradável aos

olhos. Para garantir a estabilidade temporal na conversão de vídeo 2D a 3D fazemos a propagação do mapa de profundidade. A propagação do mapa de profundidade consiste em criar um mapa de profundidade para quadro da cena do vídeo, que será herdado pelo seguinte quadro.

Herança de mapa de profundidade

A função de propagação temporal refere-se à estimação de profundidade atual a partir de uma profundidade anterior. A fim de explorar uma função de propagação temporal consistente para os objetos em movimento, se reconstruí o mapa de profundidade herdando o mapa de profundidade anterior utilizando técnicas de estimação de movimento. Com isso, evita-se a cintilação causado pela propagação de erros no mapa de profundidade. Para herdar o mapa de profundidade de um quadro para o próximo tem-se duas etapas. A primeira etapa é a exclusão dos jogadores do campo de jogo e a segunda etapa é a estimação do movimento para determinar a região e os parâmetros a ser herdados.

- **Exclusão dos jogadores do campo de jogo**

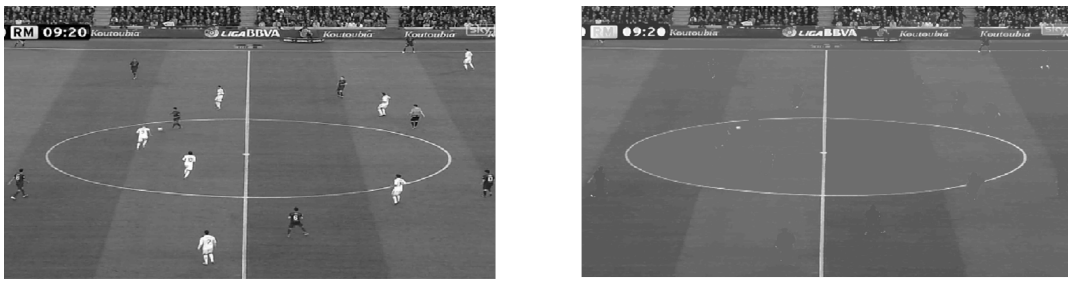


Figura 3.11: Geração da imagem do campo excluindo os jogadores (imagem de fundo): (a)imagem original; (b) imagem resultado.

Neste processo trabalha-se com imagens em escala de cinza. Para obter o resultado que se mostra na Figura 3.11-(b) a partir da imagem da 3.11-(a) (quadro original) primeiramente, é necessário atribuir o valor 0 para os *pixels* nas posições que foram detectados como jogador no processo da segmentação (Figura 3.9), obtendo um quadro com buracos como ilustra Figura 3.12. Em seguida é realizado o preenchimento de buracos simples como foi explicado na Subseção 2.4.2.

Figura 3.12: Imagem com Buracos (os buracos são *pixels* detectados como jogadores).

- **Estimação de Movimento**
  Estimação de movimento é o processo realizado para tentar encontrar os movimentos que ocorrem entre dois quadros. Esse processo consiste em procurar o lugar mais provável onde um bloco presente no quadro atual se localiza em outro quadro de referência. Estes algoritmos tem um custo computacional elevado.

  Para reduzir significativamente o custo computacional, consideramos apenas um bloco representativo (bloco central) do quadro atual e uma janela de busca no quadro de referência para realizar a estimação de movimento como ilustra a Figura 3.13. O bloco central é de tamanho de 20% da altura e 30% do comprimento do quadro e a janela de busca tem um maior tamanho em 32 *pixels* da altura e comprimento do bloco central, pois normalmente os movimentos não são muito grandes entre dois quadros seguidos (tipicamente 1/30 seg.). Também seria computacionalmente exaustivo procurar por todo o quadro.

  O critério a ser utilizado para medir a semelhança (ou diferença) é a métrica de erro médio quadrático, MSE(*Mean Squared error*) definida por:

$$MSE = \frac{1}{M} \sum_i \sum_j (P_{ij} - Pr_{ij}(x, y))^2 \tag{3.10}$$

  onde P_{ij} é um *pixel* do bloco do quadro atual, $Pr_{ij}(x, y)$ é um *pixel* do bloco no quadro de referência (previamente codificado e reconstruído) candidato a predição (deslocado em x e y) e M é o número de *pixels* pertencentes ao bloco. O bloco escolhido é o que minimiza a MSE escolhida.

  A estimação de movimento deve gerar um vetor de movimento indicando a posição de um bloco no quadro de referência, uma vez obtido o vetor $V(x_1, y_1)$ herdamos o mapa de profundidade como segue:

  - Se o movimento foi somente no sentido horizontal ($x_1 = 0$ e $y_1 \neq 0$), o mapa de profundidade é herdado sem nenhuma modificação.

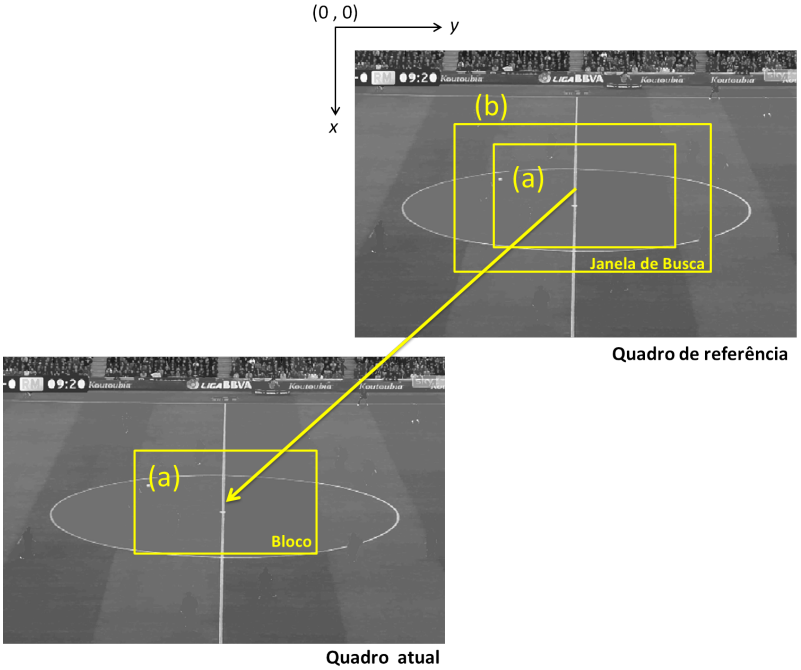


Figura 3.13: Estimação de movimento: (a) bloco do quadro atual; (b) janela de busca do quadro anterior.

– Se o movimento foi somente no sentido vertical ($x_1 \neq 0$ e $y_1 = 0$), verificamos o valor x_1. Se $x_1 > 0$ quer dizer que o movimento foi no sentido vertical descendente (a imagem está se aproximando). Desta forma x_1 linhas da parte de acima da imagem são removidas e x_1 linhas são adicionadas na parte de abaixo da imagem. A área que não foi eliminada é herdada como se mostra na Figura 3.14-(b). Para preencher as x_1 linhas que foram adicionadas com informação de profundidade utilizamos a Equação 3.9, para este caso o parâmetro $Nivel_{inicial}$ é atualizado com a equação seguinte:

$$Nivel_{inicial} = x_1 \left\lfloor \frac{Nniveis}{altura} \right\rfloor + Nivel_{inicial} \tag{3.11}$$

  Se x_1 for negativo quer dizer que o movimento foi no sentido vertical ascendente (a imagem está se afastando). Então $|x_1|$ linhas da parte de abaixo da imagem são removidos e $|x_1|$ linhas são adicionadas na parte de acima da imagem. A área que não foi eliminada é herdada como se mostra na Figura 3.14-(a). Para preencher as $|x_1|$ linhas que foram adicionadas com informação de profundidade utilizamos a Equação 3.9, para este caso o parâmetro $Nivel_{inicial}$ é atualizado com a Equaçao 3.11, tendo em conta que x_1 é negativo.

– Se o movimento for no sentido horizontal e vertical ($x \neq 0$ e $y \neq 0$), só tomamos o sentido vertical e fazemos o mesmo procedimento que o item anterior.

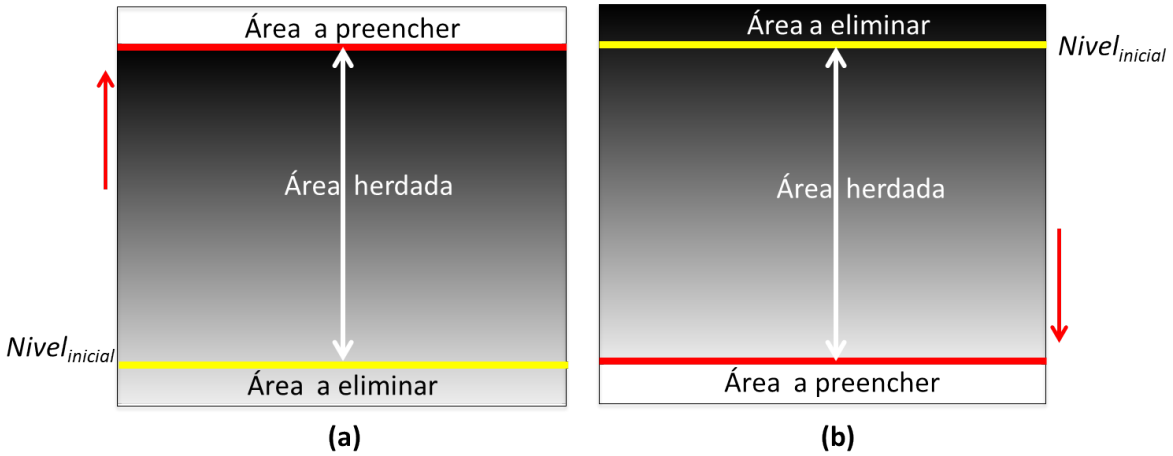


Figura 3.14: Propagação do mapa de profundidade: (a) movimento ascendente, da linha amarela para baixo é removido do mapa de profundidade e acima da linha vermelha é preenchido com a informação de profundidade (b) movimento descendente, da linha amarela para acima é removido do mapa de profundidade e abaixo da linha vermelha é preenchido com a informação de profundidade.

3.3.2 Atribuir Jogadores no Mapa de Profundidade

Uma vez obtido o mapa de profundidade, os valores de profundidade para jogadores têm que ser atribuídos. Isto é feito assumindo-se que a câmara esta alinhada verticalmente e que os jogadores estão em estreito contato com o solo.

Os jogadores então são modelados como cartazes (*billboards*) cuja profundidade é atribuída no ponto mais baixo do mapa de profundidade da região segmentada como ilustrado na Figura 3.15

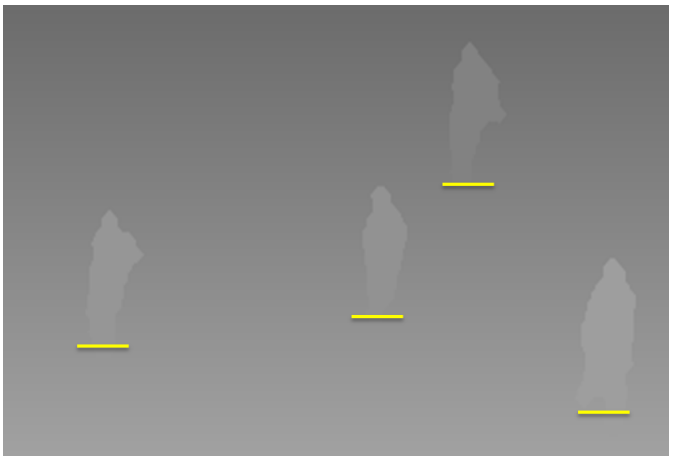


Figura 3.15: Jogadores como cartazes: a profundidade do jogador é atribuído a partir de mapa de profundidade.

3.4 Renderização estéreo

Para a renderização estéreo é utilizado o método DIBR (*Depth Image Based Rendering*), A DIRB baseia-se em apenas uma sequência de vídeo e uma sequência de mapas de profundidade correspondentes, com o qual atribui a cada *pixel* da sequência do vídeo um valor de profundidade (distância do objeto com a câmera). A partir da informação a DIBR é capaz de renderizar (quase) vistas "virtuais". Com esta técnica, sequências de vídeo para o olho esquerdo e direito podem ser processados e utilizados em vídeo 3D. A grande vantagem é que os parâmetros relevantes não são pré-definidos no momento da gravação, mas o valor de paralaxe pode ser configurado pelo receptor. Isso permite que o usuário configure a percepção de profundidade de acordo a suas preferências individuais.

DIBR toma os *pixels* da vista original e constrói um modelo simples em 3D da cena, mediante a projeção dos *pixels* no mundo 3D de acordo com seus valores de profundidade especificadas na sequência de mapas de profundidade. Este modelo 3D então é projetada sobre o plano da imagem de uma câmera virtual. Este processo é chamado de 3D *image warping*.

Ao utilizar a DIBR para vídeo 3D, os pontos de vista virtuais geralmente são deslocadas horizontalmente em relação à exibição original. Isso permite que a visão virtual de alguns objetos do primeiro plano ocluem objetos do fundo na exibição original. O fluxo de vídeo original não fornece nenhuma informação sobre os valores de cor dessas áreas de fundo, então ocorrem falhas na vista renderizada. O preenchimento das regiões não ocluídas ou "falhas de desoclusão" normalmente não pode ser feito com a informação original pois simplesmente não se encontra disponível. Para preencher as falhas com informação realiza-se processos de preenchimento de buracos sofisticados.

Após a renderização com o método DIBR tem-se como resultado imagens em formato lado-a-lado (Figura 3.16) ou anaglifo.

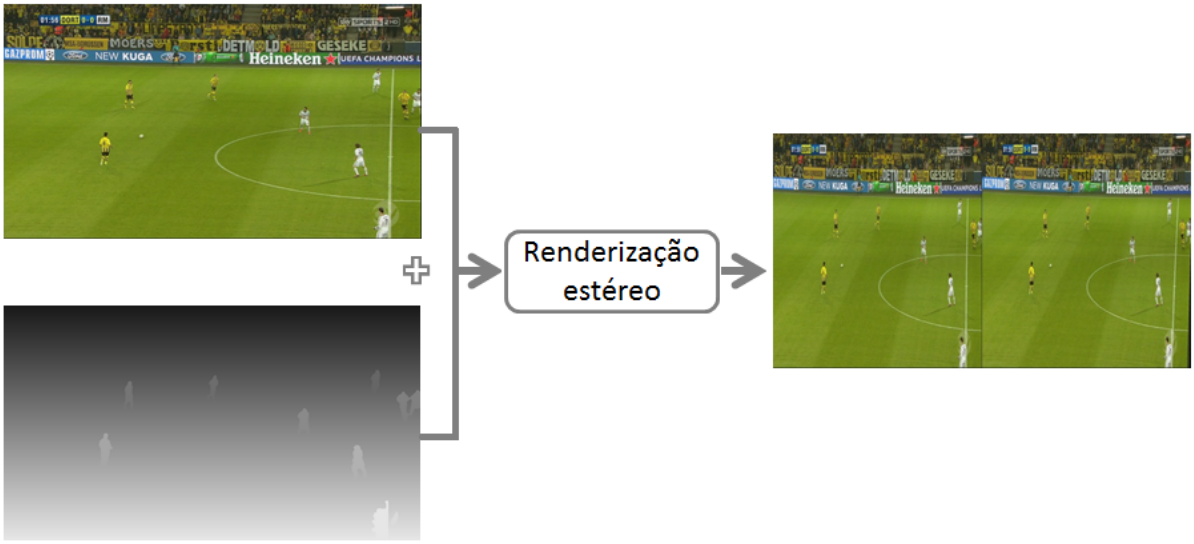


Figura 3.16: Renderização estéreo.

Capítulo 4

Resultados

4.1 Recursos utilizados

Neste dissertação utilizamos os seguintes recursos computacionais.

Software

- *MatLab* para geração dos mapas de profundidade.
- *3DCombine* para a renderização estéreo foi utilizado.
- NVIDA 3D *Vision Player* - V1.6.4, para a visualização dos vídeos 3D.
- *Stereoscopic Player* - V1.6.6

Hardware

- Processador Intel Core i7 CPU 2.80 GHz.
- 6GB de RAM
- Monitor Samsung LCD 3D 120Hz.
- Placa de Vídeo NVIDIA GeForce GTX 470
- Óculos sem fio
- Emissor infravermelho.
- Sistema Operacional: Windows 7 Ultimate 64 bits.

4.2 Resultados Experimentais

Nesta Seção é apresentada uma avaliação do processo de conversão de vídeo 2D para 3D. O conjunto de teste preparado consiste em um material de vídeo com iluminação não uniforme do campo de jogo e com vários jogadores.

Para testar o algoritmo desenvolvido ao longo do projeto, foram utilizadas sequências de 300 quadros de 1280x720 *pixels* de vídeos de futebol. Estas sequências de vídeo são de vista panorâmica com movimentos somente no campo de jogo.

Os limiares utilizados para os experimentos são mostrados na Tabela 4.1

Tabela 4.1: Limiares usados nos experimentos.

Limiar	Valor	Uso do limiar
θ	0.1	Classificador de *pixels* do campo.
$Nbins$	128	Realizar histogramas
$Nniveis$	335	Criação de mapa de profundidade.

Os resultados dos testes esta fornecido no link[1], e incentivamos os leitores a ver os resultados de vídeo em uma tela 3D de qualidade, sempre que possível.

A Figura 4.1 apresenta os diferentes cenários do campo de jogo utilizando as sequencias Vídeo-1, Vídeo-2 e Vídeo-3 nas quais foi testado o algoritmo.

Figura 4.1: Diferentes cenas obtidas aleatoriamente a partir do campo de jogo das sequências Vídeo-1, Vídeo-2 e Vídeo-3.

Apresentamos exemplos pictóricos dos resultados do Vídeo-1, Vídeo-2 e Vídeo-3 nas Figuras 4.2, 4.3 e 4.4 respectivamente, onde as imagens da primeira linha são os quadros dos vídeos de entrada, as imagens da segunda linha são mapas de profundidade

[1] https://www.dropbox.com/sh/72zb04x9udxtlih/W7tyOebvbB

correspondentes e imagens da terceira fila mostram imagens em formato anaglifo vermelho/azul. Finalmente, a quarta linha mostra imagens em formato SBS (*side-by-side* ou em português lado-a-lado). Estos dois formatos, anáglifo e SBS, são sintetizados pelos pares estereoscópicos obtidos pelos quadros de vídeo de entrada, e mapas de profundidade correspondentes.

Nos mapas de profundidade resultantes pode-se observar poucos quadros com falsos positivos e pequenas linhas do campo que formam parte dos jogadores. No entanto não são muito perceptíveis estes erros no vídeo resultante.

Com a propagação do mapa de profundidade os vídeos resultantes apresentam pouco cintilamento.

Na Figura 4.2 mostramos os resultados do Vídeo-1 para 3 cenários diferentes.

Figura 4.2: A primeira linha apresenta as imagens de entrada do Video-1, a segunda linha apresenta os mapas de profundidade, e a terceira e quarta linha apresentam as imagens estereoscópicas em formatos anáglifo e SBS.

Figura 4.3: A primeira linha apresenta as imagens de entrada do Video-2, a segunda linha apresenta os mapas de profundidade, e a terceira e quarta linha apresentam as imagens estereoscópicas em formatos anáglifo e SBS.

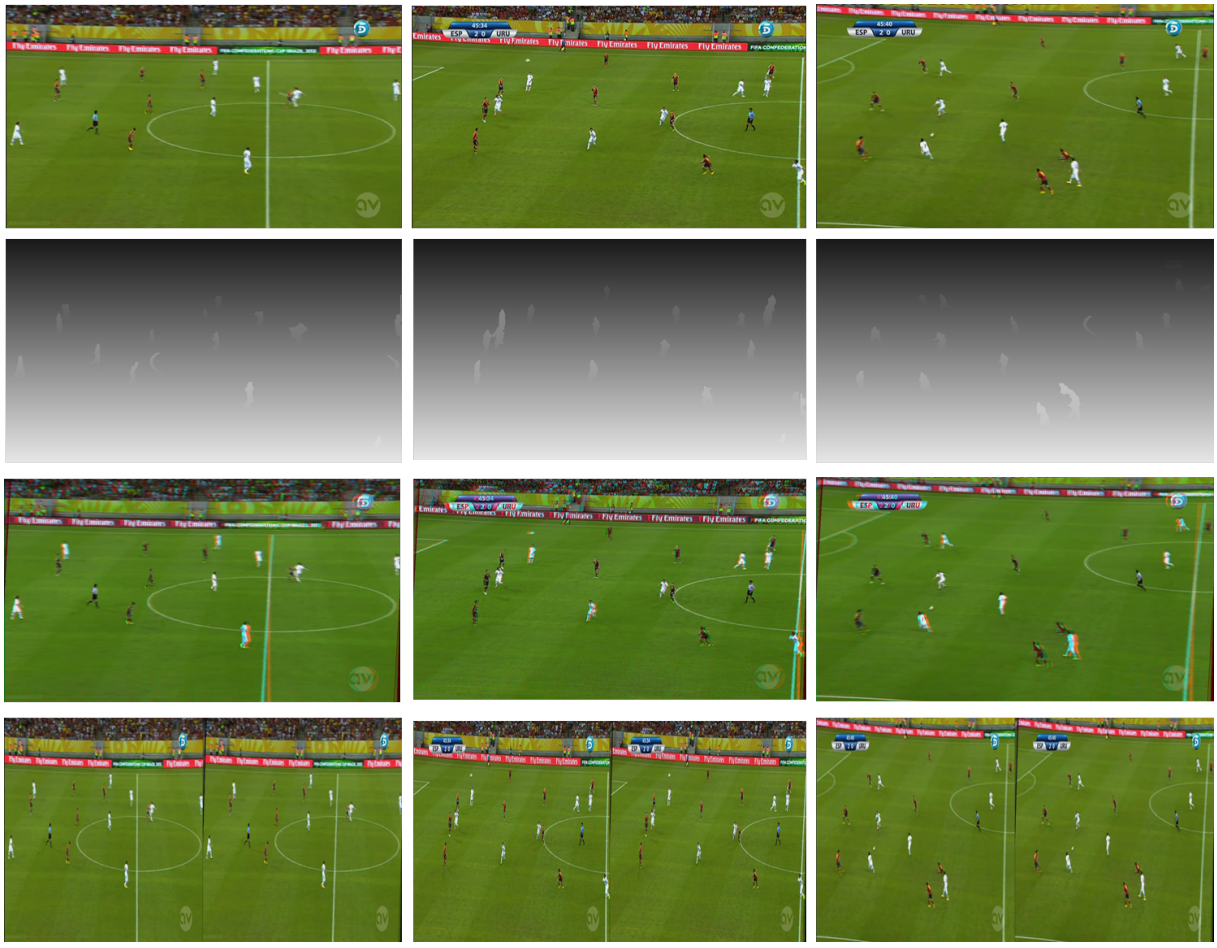


Figura 4.4: A primeira linha apresenta as imagens de entrada do Video-3, a segunda linha apresenta os mapas de profundidade, e a terceira e quarta linha apresentam as imagens estereoscópicas em formatos anáglifo e SBS.

Capítulo 5

Conclusões

Neste trabalho, apresentamos um método para criar imagens estereoscópicas de entrada monoscópica de filmagens de futebol com vistas panorâmicas. O fundo estático é tratado separadamente dos jogadores em movimento. Para o fundo estático propomos a propagação do mapa de profundidade para garantir a estabilidade temporal, que era a principal causa de artefatos visuais em métodos de esta natureza. Jogadores em movimento são tratados como cartazes ou *billboards*(objetos sem profundidade). A principal contribuição desta dissertação é:

- É realizado métodos simples para eliminar o público e as linhas do campo de jogo, após a segmentação baseado num modelo de histogramas.
- É feita uma propagação do mapa de profundidade utilizando apenas a estimação de movimento em um bloco, desta forma garante a estabilidade temporal evitando o cintilamento e diminuindo a carga computacional.

Um método de segmentação baseado num modelo de histograma de cores RGB que aprende a detectar os *pixels* do campo e agrupá-los em uma região de campo de jogo, seguido de um filtro de Sobel para eliminar a maioria das linhas do campo. Os resultados com estes simples métodos mostram em seu maioria resultados razoáveis. Embora hajam limitações no método de segmentação, já que é realizado a segmentação em um único quadro, e não são usados informações de rastreamento em sequências de imagens.

O método é simples, robusto, e pode proporcionar uma redução significativa de custos da taxa de *bits* pelo uso do formato 2D+profundidade na criação de conteúdo esportivo em 3D estereoscópico para visualização domestica. No entanto, apesar de sua simplicidade, descobrimos que o nosso método tem uma precisão suficiente para muitos casos. Isto se deve parcialmente graças à robustez na percepção estéreo humana, onde outras pistas, como a movimento paralaxe, iluminação e tamanho do objeto conhecido compensam pequenas imprecisões do processo estereoscópico.

Como trabalhos futuros podemos sugerir:

- Modificar e melhorar o método de segmentação já que é a principal fonte de artefatos em nossos resultados, e um dos principais tópicos no processo de conversão estereoscópica.
- Melhorar métodos para separar os jogadores, quando eles se sobrepõem ou ocluem parcialmente uns aos outros.

- Gerar mapas de profundidade para vistas que não sejam panorâmicas.
- Gerar mapas de profundidade para outros tipos de esportes.
- Elaborar um algoritmo de renderização estéreo em formato 2D+profundidade para transmissão em tempo real.

Referências

[1] M. O. de Beeck and A. Redert, "Three dimensional video for the home.," in *Proc. EUROIMAGE International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging (ICAV3D'01), Mykonos, Greece*, pp. 188–191, 2001. 1

[2] L. Schnyder, O. Wang, and A. Smolic, "2d to 3d conversion of sports content using panoramas," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pp. 1961–1964, 2011. 2, 25, 29, 30, 31

[3] W. Zou, "Developing end-to-end standards for 3d tv to the home," vol. 119, pp. 32–38, SMPTE Mot. Imag. J, Outubro 2010. 2

[4] A. Costa, *Compreender o Cinema*. Globo, 1989. 3

[5] R. Bhola, *Binocular vision*. The University of Iowa, Department of Ophthalmology Visual and Sciences, January 2006. 4

[6] "Estereoscopia imagens 3d cruzando os olhos." http://www.adrenaline.uol.com.br/forum/geral/243526-estereoscopia-imagens-3d-cruzando-os-olhos.html/. Acessado em Março, 2012.

[7] B. Bowers, I. of Electrical Engineers, and S. Museum, *Sir Charles Wheatstone FRS: 1802-1875*. History of technology series, Institution of Electrical Engineers, 2001. 4

[8] O. W. Holmes, "The american stereoscope," vol. 1, (New York), Journal of Photography of the George Eastman House, Março. 5

[9] H. Morgan and D. Symmes, *Increíble 3-D*. Boston, Nueva York: Little, Brown and Company, 1982. 5

[10] "Stereo cameras 3dham." http://www.3dham.com/stereo_cameras/index.html. Acessado em Julho, 2013.

[11] J. Glancey, "Classics of everyday design no 48." http://www.guardian.co.uk/artanddesign/2008/jul/31/viewmaster.design.classic. Acessado em Julho, 2013.

[12] "History of imax corporation." http://www.imax.com/corporate/history/. Acessado em Julho, 2013. 6

[13] D. Isbell and F. Donnell, *Mars Pathfinder Landing*. NASA, Julho 1997. 6

[14] D. Boesten and P. Vandewalle, *Depth estimation for stereo image pairs.* Pearson Education, 2010. 7

[15] C. Chinnock, *The state of 3d in the home.* Insight Media for the 3D Home Consortium, Abril 2010. 10

[16] *3D Video - Multiple systems being developed.* Victor Company of Japan, 2009. 11

[17] A. J. Woods and C. R. Harris, "Comparing levels of crosstalk with red/cyan, blue/yellow, and green/magenta anaglyph 3d glasses," in *Proceedings of SPIE Stereoscopic Displays and Applications XXI*, vol. 7253, pp. 0Q1–0Q12, Janeiro 2010. 12

[18] M. K. Gandhi, "How 3D Works." http://www.onlineschools.org/blog/how-3d-works/. Acessado em Março, 2012. 15

[19] C. Laburú, A. Simões, and A. Urbano, "Caderno catarinense de ensino de física," in *Departamento de Física, Universidade Federal de Santa Catarina*, vol. 15, (Florianópolis, Brasil), 1998. 14

[20] J. L. Prados Ribeiro and M. d. F. Da Silva Verdeaux, "Reflexão e polarização em óculos 3d," in *Física na Escola*, vol. 13, 2012. 14

[21] *Estado del Arte de las Tecnologías audiovisuales.* Xpertia Soluciones Integrales, Cluster ICT-Audiovisual de Madrid, 2012. 16

[22] S. Battiato, A. Capra, S. Curti, and M. La Cascia, "3d stereoscopic image pairs by depth-map generation," in *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 3DPVT.*, pp. 124–131, 2004. 17, 39

[23] K. S. Fu and J. K. Mui, "A survey on image segmentation," in *Pattern Recognition*, vol. 13, pp. 3–16, 1981. 17

[24] R. C. Gonzalez and R. E. Woods, *Digital Image Processing.* Upper Saddle River, NJ: Prentice-Hall, Second Ed., 2002. 18, 19, 36

[25] J. C. Russ, *The image processing handbook.* CRC Press, 2007. 19

[26] L. Zhang, G. Dugas-Phocion, J.-S. Samson, and S. M. Seitz, *Single View Modeling of Free-Form Scenes.* IEEE Pattern Analysis and Machine Intelligence, Março 2001. 24, 25, 26

[27] A. van den Hengel, A. Dick, and Thorm 24, 26, 27

[28] M. S. A. Saxena and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 824–840, Março 2009. 24, 26, 27

[29] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," in *ACM Transactions on Graphics (TOG)*, vol. 24, (New York, NY, USA), pp. 577–584, ACM, 2005. 24, 27, 28

[30] C. Hou, J. Yang, and Z. Zhang, "Stereo image displaying based on both physiological and psychological stereoscopy from single image.," in *ACM International Journal of Imaging Systems and Technology - Multimedia*, vol. 18, (New York, NY, USA), pp. 146–149, John Wiley & Sons, Inc., Agosto 2008. 25, 28

[31] S. Gortler and M. Cohen, *Variational modeling with wavelets*. Princeton Univ: Dept of Computer Science, 1994. 25

[32] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications*. American Mathematical Society, 1980. 26

[33] J. M. Loomis, *Looking down is looking up*. Nature News and Views, Novembro 2001. 28

[34] A. P. Pentland, "A new sense for depth of field," vol. PAMI-9, pp. 523–531, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1987. 29

[35] Y. Huang, J. Llach, and S. Bhagavathy, "Players and ball detection in soccer videos based on color segmentation and shape analysis," in *Proceedings of the 2007 international conference on Multimedia content analysis and mining*, MCAM'07, (Berlin, Heidelberg), pp. 416–425, Springer-Verlag, 2007. 31, 32

[36] J. R. Wang and N. Parameswaran, *Survey of Sports Video Analysis: research issues and applications*. Proceedings of the Pan-Sydney Area Workshop on Visual Information Processing VIP03 Pages 87-90, 2003. 32

[37] M. Jones and J. Rehg, "Statistical color models with application to skin detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 274–280, 1999. 32