UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
COORDENAÇÃO DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Dissertação de Mestrado

Um Mecanismo de Atenção Visual Integrando Evidências Espaciais e Temporais

Sandberg Marcel Santos

Campina Grande – PB
Agosto de 2005

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
COORDENAÇÃO DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Um Mecanismo de Atenção Visual Integrando Evidências Espaciais e Temporais

Sandberg Marcel Santos

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática da Universidade Federal de Campina Grande como parte dos requisitos necessários para obtenção do grau de Mestre em Informática.

Orientadores:
Herman Martins Gomes, PhD
Díbio Leandro Borges, PhD

Linha de Pesquisa: Modelos Computacionais e Cognitivos

Campina Grande – PB
Agosto de 2005

SANTOS, Sandberg Marcel

S231M

Um Mecanismo de Atenção Visual Integrando Evidências Espaciais e Temporais.

Dissertação (Mestrado), Universidade Federal de Campina Grande, Centro de Ciências e Tecnologia, Coordenação de Pós-Graduação em Informática, Campina Grande – Paraíba, Agosto de 2005.

101p.Il

Orientadores: Herman Martins Gomes
Díbio Leandro Borges

Palavras- Chave:

1. Inteligência Artificial
2. Atenção Visual Espacial
3. Atenção Visual Temporal
4. Integração de Evidências
5. Detecção de Transições
6. Segmentação de Objetos em Vídeo

CDU – 007.52

"UM MECANISMO DE ATENÇÃO VISUAL INTEGRANDO EVIDÊNCIAS ESPACIAIS E TEMPORAIS"

SANDBERG MARCEL SANTOS

DISSERTAÇÃO APROVADA EM 29.08.2005

PROF. HERMAN MARTINS GOMES, Ph.D
Orientador

PROF. DÍBIO LEANDRO BORGES, Ph.D
Orientador

PROFª JOSEANA MACÊDO FECHINE , D.Sc
Examinadora

LUIZ MARCOS GARCIA GONÇALVES, Dr.
Examinador

CAMPINA GRANDE – PB

Agradecimentos

Agradeço primeiramente e acima de tudo a Deus, a quem tudo devemos e em quem tudo somos.

Agradeço ao Professor Herman Martins Gomes, pela orientação, paciência e dedicação dispensadas.

Agradeço ao Professor Díbio Leandro Borges, pela participação e ajuda neste trabalho.

Agradeço a toda a equipe da Copin: professores, secretárias e demais funcionários, cujo trabalho é essencial para o andamento normal das atividades de todos.

Agradeço aos meus pais, José Arnaldo e Edirlene, pela educação que me deram e pela participação e auxílio constantes em minha vida, apesar da distância.

Agradeço à minha noiva, Loraine, pela companhia, carinho e ajuda em todos os momentos.

Por fim, agradeço a todos os familiares, amigos, colegas e a todos aqueles que, direta ou indiretamente, contribuíram para a conclusão deste trabalho.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Resumo

Sistemas visuais biológicos utilizam uma estratégia que prioriza a extração das informações mais relevantes à execução de uma determinada tarefa visual, reduzindo assim a quantidade de recursos computacionais necessários para realizá-la. Tomando como inspiração essa estratégia, tem-se a Atenção Visual, área da Visão Computacional que se preocupa com o processamento de cenas visuais, procurando encontrar as regiões mais salientes (mais importantes de serem analisadas). Nesse contexto, o presente trabalho propõe um novo modelo de Atenção Visual que integra diferentes abordagens: Atenção Espacial (ou Estática) bottom-up, Atenção Temporal (ou Dinâmica) e Visão Estéreo. Este trabalho também desenvolve, para as duas primeiras abordagens, uma implementação que é validada através de uma série de experimentos. Apresentam-se uma estratégia para se realizar a segmentação de objetos móveis em cenas visuais, como parte integrante do modelo proposto, e um estudo de caso envolvendo a utilização das evidências temporais, obtidas pelo sistema de Atenção Visual desenvolvido, no problema de detecção de transições abruptas em seqüências de vídeo. Os resultados obtidos indicaram que a estratégia proposta para a extração de características temporais e para a detecção de objetos móveis se constituiu em uma forma simples e versátil para se realizar a detecção e a segmentação de movimento em vídeos. Já no que se refere aos experimentos envolvendo a detecção de transições abruptas, foi realizada uma avaliação de desempenho, na qual foram observadas taxas de erro reduzidas. Finalmente, a integração de características espaciais no contexto acima resultou em uma estratégia interessante para se combinar evidências das Atenções Espacial e Temporal.

Abstract

Biological vision systems have mechanisms that focus on the extraction of the most relevant information for performing a given visual task, so that the overall computational effort is reduced. Inspired by these mechanisms, Visual Attention emerges as the area of Computer Vision that is mainly concerned with the processing of visual scenes, searching for the most salient regions (which are the most important to be analyzed). Within this context, the present work proposes a new Visual Attention model, which integrates different approaches: Spatial (or Static) bottom-up Attention, Temporal (or Dynamic) Attention and Stereo Vision. This work also develops, for the first two approaches, an implementation that is validated through a series of experiments. Furthermore, this work presents a strategy for the segmentation of moving objects in visual scenes, as part of the proposed model, and a case study, involving the utilization of temporal evidences from the developed Visual Attention system in the problem of detecting sharp transitions in video sequences. The results have shown that the strategy proposed for the temporal feature extraction and for the detection of moving objects was a simple and versatile way to perform motion detection and segmentation in videos. With regards to the experiments involving the detection of sharp video transitions, a performance evaluation revealed low error rates. Finally, the integration of spatial features into the above context yielded an interesting approach to combine evidence from both Spatial and Temporal Visual Attention.

Capítulo 1

Introdução

A visão é o mais complexo dos sentidos do ser humano e realiza um processo perceptivo centrado em objetivos [Machado, 1994]. Quando estimulado, o sistema visual utiliza uma estratégia baseada na economia de esforço, dando prioridade à extração das informações necessárias à execução de uma determinada tarefa. Os seres vivos classificados em posições mais baixas na escala evolutiva fazem um uso simplista da visão, voltado basicamente para a sobrevivência [Marr, 1982]. Já nos seres que ocupam posições mais altas, a visão desempenha um papel mais importante, preservando, entretanto, a característica de ser orientada por objetivos.

Por ser uma tarefa de representação e processamento de informações, a visão pode ser tratada computacionalmente [Marr, 1982]. Contudo, apesar das grandes descobertas feitas nos últimos 50 anos, uma série de aspectos da visão ainda é desconhecida para a ciência, não se sabendo exatamente como ela ocorre em nível fisiológico e psíquico, principalmente nas camadas de mais alto nível.

No que se refere à definição de visão, percebe-se que diferentes autores o fazem de maneira semelhante. Analisando-se as citações abaixo, vê-se que, por exemplo, a visão é citada por todos como um processo de construção de descrições da realidade:

> *"A análise de imagens objetiva a construção de descrições de cenas, baseando-se em informações extraídas de imagens ou seqüências de imagens."* [Rosenfeld, 1983]

> *"Visão é um processo que produz, a partir de imagens do mundo externo, uma descrição que é útil ao observador e que não é afetada por informações irrelevantes."* [Marr, 1982]

> *"... a geração de uma descrição do mundo por meio de uma ou mais imagens."* [Horn, 1986]

> *"A Visão Computacional trata da construção de descrições significativas e explícitas de objetos físicos através de imagens."* [Ballard and Brown, 1982]

Entre os animais, a percepção de estímulos provenientes do meio ambiente se constitui como atividade essencial no que diz respeito à sobrevivência. Já entre animais que possuem um sistema visual mais desenvolvido (maior parte dos estímulos importantes detectados pela visão), como os mamíferos, o reconhecimento dos estímulos visuais desempenha um papel primordial [Gonçalves, 1999]. Nesse contexto, possui grande importância a chamada atenção visual, característica dos sistemas visuais biológicos responsável por selecionar as informações mais relevantes em cenas visuais. A atenção visual é determinante para a perpetuação e a evolução das espécies, na medida em que se caracteriza como a habilidade de fixar rapidamente a visão em pontos de interesse e reconhecer possíveis presas, predadores ou rivais [Itti and Koch, 2001].

Para se realizar a percepção dos estímulos visuais em máquinas, utiliza-se um conjunto de métodos e técnicas de análise e interpretação de imagens conhecido como Visão Computacional. No contexto da Visão Computacional, tem-se a área de Atenção Visual Computacional (análoga à atenção visual biológica), que se dedica a desenvolver mecanismos que reduzam o custo computacional quando do processamento de cenas visuais em máquinas. A partir deste ponto, a expressão Atenção Visual será utilizada para referenciar as versões computacional e biológica indistintamente. Faz-se importante ressaltar, entretanto, o fato de que, no meio científico, no que se refere à atenção visual biológica, apenas se possui evidências de seu funcionamento, e que foi exatamente a partir de tais evidências que foram convencionados mecanismos para a realização da Atenção Visual Computacional.

A Atenção Visual possui duas abordagens [Coull and Nobre, 1998], a espacial (que se preocupa em determinar estaticamente quais regiões da cena são mais relevantes) e a temporal (que se preocupa em dinamicamente determinar em que região e momento do tempo realizar uma análise mais detalhada, ou seja, quais cenas e regiões nas cenas em uma seqüência são as mais relevantes). A Atenção Espacial, por sua vez, pode ser *bottom-up* (ou atenção baseada em características primitivas) ou *top-down* (ou atenção baseada em modelos de objetos) [Itti and Koch, 2001]. A Atenção Visual será mais detalhadamente explicada no Capítulo 2. Vale frisar que essa divisão em Atenção Espacial *bottom-up*, Espacial *top-down* e Temporal é adotada do ponto de vista

computacional. Já no que se refere aos aspectos biológicos, as linhas que separam esses diferentes tipos de atenção são bem mais tênues.

São inúmeras as aplicações práticas de um sistema de Atenção Visual, dentre as quais pode-se citar: monitoramento de vídeo (*video surveillance*) [Collins et al., 2000; Wildes, 1998], renderização de vídeo [Yee et al., 2001], recorte automático de vídeo [Wang et al., 2004; Chen et al., 2003], detecção de transições [Guimarães et al., 2001], problemas de indexação de vídeo (em empresas de televisão), bancos de dados multimídia, detecção de faces, entre outras.

Na próxima seção, os principais objetivos do trabalho serão expostos e se discutirá sua relevância para as áreas de Visão Computacional e Atenção Visual. Já na segunda seção, explica-se como este trabalho está estruturado.

1.1 Trabalhos Realizados e Relevância

Os dois principais objetivos deste trabalho foram (i) a proposta e a implementação de um modelo de Atenção Visual (exposto no Capítulo 4) combinando as abordagens espacial *bottom-up* (Seção 2.1) e temporal (Seção 2.2), e (ii) o desenvolvimento de experimentos de segmentação de objetos móveis em cenas visuais (Seção 5.3) e de detecção de transições abruptas em seqüências de vídeo (Seção 5.2), utilizando as evidências temporais obtidas a partir do sistema de Atenção Visual desenvolvido.

A proposta do modelo de Atenção Visual se concretizou em um modelo contendo os seguintes módulos: um Módulo de Atenção Temporal (que extrai, a partir dos quadros do vídeo original, evidências sobre o movimento presente nas cenas visuais), um Módulo de Visão Estéreo (que obtém evidências sobre a profundidade das cenas visuais, a partir dos quadros dos vídeos obtidos de diferentes câmeras, posicionadas em diferentes posições e capturando as mesmas cenas), um Módulo de Segmentação de Movimento e de Profundidade (que gera, para cada quadro do vídeo original, um novo quadro, em que as regiões onde ocorre movimento e que estão mais próximas à câmera ficam em destaque) e um Módulo de Atenção Espacial *bottom-up* (que extrai as evidências *bottom-up* dos quadros gerados pelo Módulo de Segmentação, efetivando a integração entre as Atenções Espacial e Temporal).

Já no que se refere à implementação desse modelo de Atenção Visual, não foi realizada, por questões de escopo, a implementação do Módulo de Visão Estéreo, uma vez que algoritmos de Visão Estéreo possuem, normalmente, alta complexidade e requerem um processo de calibração muito sensível a mudanças nos parâmetros de aquisição. Assim sendo, a integração de evidências de Visão Estéreo foi deixada como um trabalho futuro. Para o Módulo de Atenção Espacial *bottom-up*, por sua vez, foi utilizado um programa desenvolvido no trabalho de Rodrigues [Rodrigues, 2002] e gentilmente cedido pelo autor. Quanto aos demais módulos, eles foram implementados e testados segundo suas descrições no modelo, sendo feita a observação de que o Módulo de Segmentação foi implementado apenas como um Módulo de Segmentação de Movimento, não agregando nenhuma informação referente à profundidade das cenas visuais.

Os experimentos em segmentação de objetos móveis foram concretizados a partir do Módulo de Segmentação de Movimento. Vale frisar também que, nos experimentos em segmentação, foi incluída a integração entre as evidências espaciais (*bottom-up*) e temporais, por meio do processamento da saída do Módulo de Segmentação pelo Módulo de Atenção Espacial *bottom-up*. Por fim, utilizando-se as evidências de movimento fornecidas pelo Módulo de Atenção Temporal, foram realizados experimentos em detecção de transições abruptas em vídeos, sendo estes os experimentos explorados em maiores detalhes, dentre todos os experimentos realizados neste trabalho.

Passando ao tema da relevância deste trabalho para a área de pesquisa, tem-se que a área de Atenção Visual vem despertando interesse crescente nos últimos anos. Em particular, a Atenção Visual Espacial *bottom-up* já foi bastante estudada e, até mesmo a *top-down*, que inicialmente não era muito explorada, foi coberta em maior profundidade, no início da década corrente, por trabalhos como os desenvolvidos por Sun e Fisher, e por Itti e Koch [Sun and Fisher, 2003; Sun and Fisher, 2002; Itti and Koch, 2001; Itti, 2000]. Entretanto, no que diz respeito à Atenção Temporal, há uma grande margem para a pesquisa. Trata-se de uma área com avanços limitados nas últimas décadas, devido ao enorme volume de dados que precisa ser tratado pelos algoritmos. No entanto, com os aumentos recentes da velocidade de processamento e da capacidade de memória, obtiveram-se melhores recursos computacionais para pesquisas

em novos algoritmos. Na Figura 1, é apresentado um gráfico comparativo, relacionando os tipos de Atenção Visual e suas respectivas complexidades computacionais.

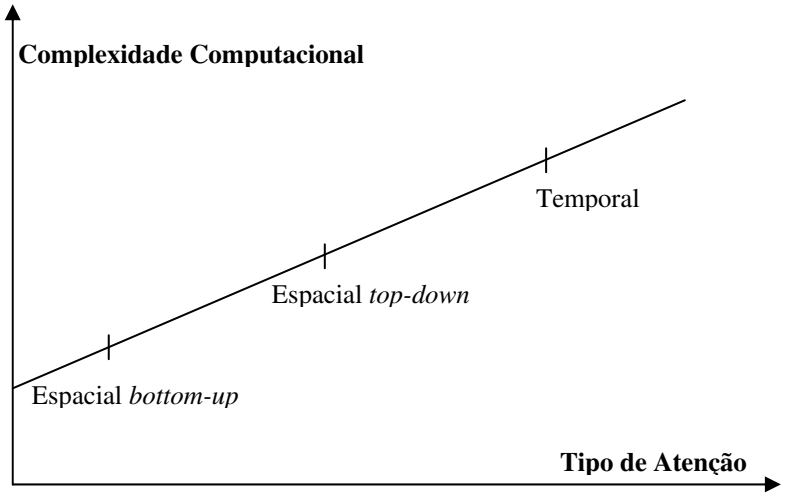


Figura 1 – Tipo de Atenção x Complexidade Computacional

Como se pode ver pelo gráfico da Figura 1, a Atenção Espacial *bottom-up* tem baixa complexidade, pois é independente de conteúdo e realiza análise local de *pixels* em um único quadro de vídeo. Já a Atenção Espacial *top-down* apresenta uma complexidade intermediária, por depender de conteúdo e realizar análise global de imagens, ainda que em um único quadro. Por fim, a Atenção Temporal possui alta complexidade, podendo ser dependente ou independente de conteúdo e realizando análise tanto local quanto global em vários quadros.

A Atenção Temporal dá maior ênfase à detecção de ocorrências relevantes em determinados instantes do tempo [Coull and Nobre, 1998], como, por exemplo, objetos em movimento nas cenas visuais (usualmente considerados alvos da Atenção Visual). Conforme mencionado anteriormente, a Atenção Visual (seja espacial ou temporal) visa a redução do custo computacional para se analisar imagens (em seqüência ou não). A Atenção Espacial realiza tal redução através da seleção das regiões de maior importância na imagem, tornando necessária a análise somente dessas regiões e não de toda a imagem. Integrando-se elementos temporais aos elementos espaciais, é possível aperfeiçoar a detecção de regiões-alvo da Atenção Visual quando se possui uma seqüência de imagens. Ao se analisar as imagens em seqüência, os dados obtidos de uma determinada imagem vêm a simplificar consideravelmente a análise de uma imagem posterior. É desse modo que ocorre a redução do custo computacional e é nesse

sentido que os estudos aqui feitos, visando a integração entre as Atenções Espacial e Temporal, servem como uma contribuição para a pesquisa de Atenção Visual.

Já no que se refere especificamente à Atenção Espacial, uma inovação presente no modelo proposto neste trabalho consiste na proposta de integração de evidências de profundidade a partir de um Módulo de Visão Estéreo. A Visão Estéreo se caracteriza como uma evidência de Atenção Espacial *bottom-up* e utiliza mecanismos semelhantes ao da Atenção Temporal. Como citado anteriormente, por questões de escopo, nesta dissertação, não foram realizados uma implementação nem experimentos envolvendo Visão Estéreo, tendo tal integração sido deixada como um trabalho futuro.

1.2 Estrutura da Dissertação

A partir deste ponto, o trabalho está organizado como segue. O Capítulo 2 introduz o conceito de Atenção Visual e trata dos modelos computacionais de Atenção Visual, mais especificamente Atenção Espacial (*bottom-up* e *top-down*) e Temporal, e da Visão Estéreo.

No Capítulo 3, serão discutidos os principais trabalhos relacionados às áreas de pesquisa, mais especificamente trabalhos que lidam com mecanismos biológicos da Atenção Visual e com vários aspectos das Atenções Espacial *bottom-up* e Temporal.

O Capítulo 4 trata do protótipo do sistema de Atenção Visual proposto, de sua arquitetura geral e de especificidades sobre um de seus módulos, o Módulo de Atenção Temporal.

No Capítulo 5, são exibidos experimentos realizados com os módulos do protótipo do sistema e experimentos de integração entre os módulos. Os resultados obtidos são comentados e avaliados. Além disso, são discutidas duas aplicações que se utilizam das evidências da Atenção Temporal: um experimento em segmentação de movimento e um estudo de caso para a detecção de transições abruptas em vídeos. Também se comentam detalhes de implementação.

Finalmente, o Capítulo 6 apresenta as considerações finais sobre o trabalho desenvolvido e perspectivas para trabalhos futuros.

Capítulo 2

Atenção Visual

Os primatas possuem uma grande capacidade de interpretar cenas complexas em tempo real, apesar da velocidade limitada do circuito neural disponível para uma tarefa dessa natureza. O tempo médio de disparo de um neurônio é da ordem de 10 milissegundos [Bauchspiess, 2002], o qual pode ser considerado alto quando comparado ao tempo de execução de uma instrução de máquina elementar em um processador, da ordem de um nanossegundo. Tendo-se como base estudos psicofísicos em seres humanos e eletrofisiológicos em macacos, acredita-se que os processos visuais de níveis intermediário e alto selecionam apenas um subconjunto da informação sensorial disponível [Tsotsos et al., 1995], na forma de uma região circular do campo visual, conhecida como *foco de atenção*. O poder de representação neural do mundo visual é ampliado dentro da área restrita do foco de atenção, e somente essa informação é submetida para processamentos de mais altos níveis, o que favorece uma redução na complexidade de processamento.

Existem duas abordagens para a Atenção Visual [Coull and Nobre, 1998]: uma que direciona o foco de atenção para uma determinada região de uma imagem estática (Atenção Espacial) e outra que analisa uma seqüência de imagens, procurando saber que instante de tempo é maior merecedor de uma análise mais detalhada (Atenção Temporal). A Atenção Espacial, detalhada na Seção 2.1, pode envolver análise de características ou evidências de baixo nível (*bottom-up*) e modelos de alto nível (*top-down*) [Itti and Koch, 2001]. A Atenção *bottom-up* normalmente resulta da integração de diversas características primitivas, como bordas, orientações etc, as quais são extraídas localmente da imagem de entrada. Um elemento diferencial desta dissertação é a proposta de incorporação de evidências de profundidade à Atenção *bottom-up*. Nesse sentido, discute-se, ao final deste capítulo, a noção de Visão Estéreo, a qual permite obter evidências de profundidade das superfícies visíveis, e que utiliza um mecanismo

para sua computação que é semelhante ao cálculo de Fluxo Óptico, utilizado para estimar movimento na Atenção Temporal (discutida na Seção 2.2).

2.1 Atenção Espacial

A Atenção Espacial preza por determinar, em imagens estáticas, as regiões-alvo do processo atencional. A análise da imagem é feita de duas maneiras: uma rápida, grosseira, maciçamente paralela, independente da tarefa e baseada nas saliências, conhecida como abordagem *pré-atencional* ou *bottom-up* (de baixo para cima), e outra mais lenta, mais detalhada, seqüencial, dependente da tarefa e controlada pela vontade, conhecida como abordagem *top-down* (de cima para baixo) [Itti, 2003; Niebur and Koch, 1998].

A reprodução da capacidade de detecção não-específica de alvos na Atenção *bottom-up* tem aplicações importantes, como, por exemplo, em sistemas embarcados (embutidos) para auxílio de navegação e em navegação robótica. Em um modelo desse tipo de atenção, o primeiro estágio de processamento é a computação das propriedades elementares das imagens (brilho, textura, cor, fluxo de movimento, disparidade estéreo, entre outras), conhecidas como *características visuais primitivas*. O gradiente dessas grandezas é usado para segmentar a imagem, produzindo-se vários mapas de características, que, combinados, dão origem a um *mapa de saliência*. Na Figura 2, são exibidas quatro imagens: a original e três exemplos de mapas de características gerados a partir dela (intensidade, cor vermelha e cor verde). Já na Figura 3, é exibido o mapa de saliência referente à imagem original da Figura 2.

Através da chamada *representação piramidal*, pode-se representar as características visuais primitivas de uma mesma imagem em diferentes escalas. A pirâmide, na verdade, é um conjunto de imagens, em que cada nível é uma cópia da imagem original (nível 0), com densidade e resolução menores do que a do nível anterior. O objetivo é obter amostras da imagem nas quais detalhes indesejados são suprimidos, ruídos são eliminados, características grosseiras são realçadas etc. A *Pirâmide Gaussiana* [Burt and Adelson, 1983] é a representação mais comum. As imagens dessa pirâmide são obtidas através de uma filtragem passa-baixa de convolução Gaussiana. Um filtro passa-baixa atenua as altas freqüências espaciais de uma imagem e

acentua as baixas freqüências. Na Figura 4, é exemplificada uma pirâmide Gaussiana de cinco níveis.

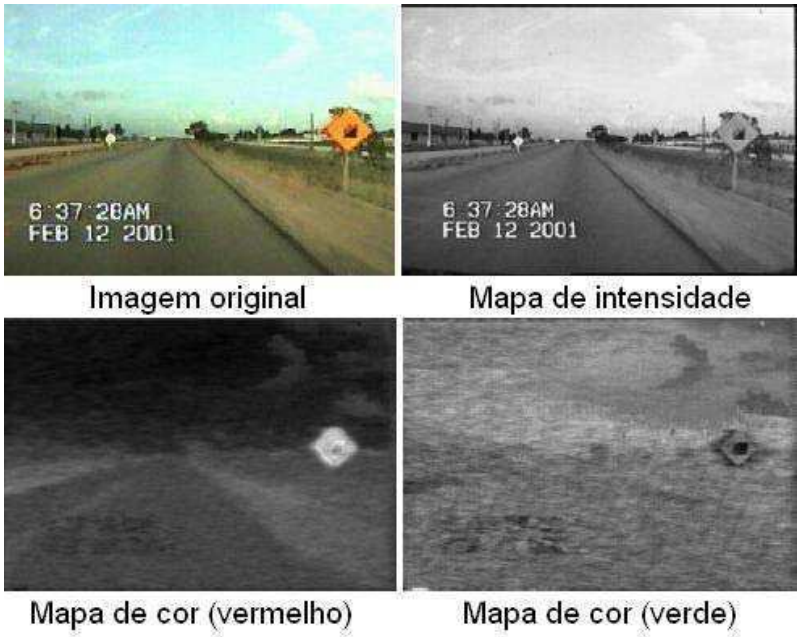


Figura 2 – Imagem original e alguns mapas de características
(extraídos de [Rodrigues, 2002])

Figura 3 – Mapa de saliência
(extraído de [Rodrigues, 2002])

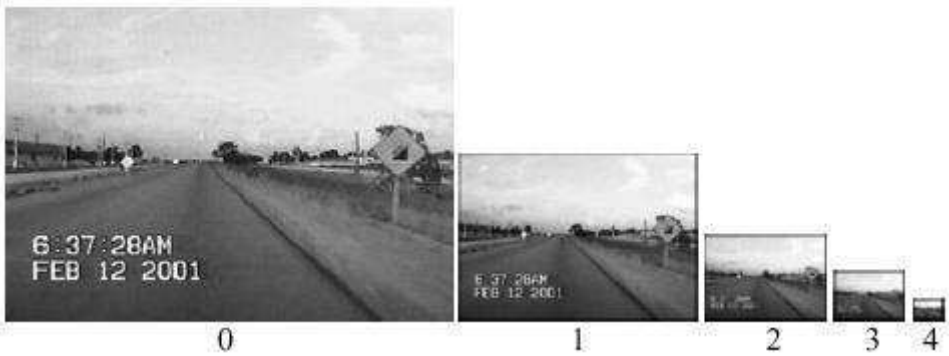


Figura 4 – Pirâmide Gaussiana
(extraída de [Rodrigues, 2002])

Enquanto a Atenção *bottom-up* é independente da natureza da tarefa e opera rapidamente (de 25 a 50 ms por item, em seres humanos [Itti and Koch, 2001]), a outra forma de Atenção Espacial, a *top-down*, é muito mais deliberativa e poderosa, possuindo um critério de seleção de variáveis, dependente da tarefa que se tem em mãos. Por ser fortemente dirigida pela vontade, a Atenção *top-down* paga um preço alto: o tempo, de 200 ms ou mais em seres humanos [Itti and Koch, 2001], equiparável ao tempo necessário para se movimentar os olhos. A Atenção *top-down* baseia-se em modelos de objetos, como esquematizado na Figura 5.

Figura 5 – Esquematização dos objetos (representados por elipses na figura) de uma abordagem de Atenção *top-down*

Ambos os processos – a atenção automaticamente atraída por características do mundo visual (Atenção *bottom-up*) e o esforço voluntário de voltar a atenção para outros objetos ou locais (Atenção *top-down*) – podem operar em paralelo. A maioria dos modelos que aplicam uma abordagem *top-down* se valem desse fato, utilizando, em verdade, uma abordagem conjunta *bottom-up/top-down*. Evidências [Itti, 2000] têm se acumulado a favor dessa abordagem combinada para o controle das regiões-alvo da Atenção Visual.

Os mecanismos exatos através dos quais ocorrem deslocamentos voluntários de atenção (Atenção *top-down*) permanecem indefinidos, apesar de alguns estudos terem reduzido o número de áreas cerebrais diretamente envolvidas [Itti and Koch, 2001] (o controle desse tipo de atenção é feito provavelmente nas áreas superiores do cérebro, inclusive nos lóbulos frontais). A seguir, são apresentados dois casos em que ocorre esse tipo de atenção: (i) atenção dirigida voluntariamente para um dentre vários

estímulos idênticos e (ii) grandes diferenças nos movimentos dos olhos dependendo da tarefa.

O ser humano é capaz de deslocar voluntariamente sua atenção dentro de seu campo visual, não importando quão uniforme seja a cena [Itti and Koch, 2001]. Um experimento típico de Atenção *top-down* consiste em exibir um conjunto de estímulos visualmente idênticos a um observador e guiá-lo, apenas em um elevado nível cognitivo (verbalmente, por exemplo), na direção de um dos estímulos. A detecção do estímulo é significantemente melhor (feita em menor tempo, por exemplo) em uma situação em que se guia o observador do que na situação oposta. Tal experimento sugere que o deslocamento voluntário da atenção na direção de um estímulo melhora a percepção desse estímulo.

Do mesmo modo, experimentos que envolvem decisões demonstram-se mais eficientes se o estímulo for detectado através de um atributo previamente conhecido. Assim, observa-se que o ser humano é capaz de selecionar voluntariamente características específicas de um estímulo. Esses resultados retratam as vantagens de se utilizar a abordagem conjunta *bottom-up/top-down*. Em suma, a Atenção *bottom-up* se concentra em características de baixo nível, focando naquilo que é visualmente mais saliente. A Atenção *top-down* fornece informações de alto nível a respeito do que está sendo procurado, restringindo a quantidade de estímulos (e/ou de características de estímulos) que devem ser analisados pelos mecanismos *bottom-up*.

Já no que diz respeito aos movimentos dos olhos, os experimentos comprovam que há uma forte influência da tarefa que se quer realizar. Tais experimentos consistem em vários observadores contemplando cenas idênticas, mas cada um com um objetivo diferente (cada um à procura de um estímulo específico pré-determinado). Verifica-se que, apesar das cenas observadas serem idênticas, os movimentos dos olhos de duas pessoas distintas são radicalmente diferentes, comprovando-se a forte influência da tarefa (influência *top-down*) no deslocamento da atenção.

2.2 Atenção Temporal

O exemplo a seguir ilustra uma modalidade de Atenção Temporal em ação, que se caracteriza por antecipar a ocorrência de um evento visual futuro a partir de evidências temporais. Imagine-se um semáforo para pedestres, cuja luz verde pisca

momentos antes do sinal fechar (tornar-se vermelho). A luz verde piscante serve como uma evidência, por meio da qual pode-se antecipar o fechamento do sinal. Focar a atenção em algo (nesse caso, no semáforo) com uma certa antecedência gera respostas mais rápidas do que na situação em que sua mente está distraída com alguma outra coisa.

Como visto na seção anterior, a Atenção Espacial tem sido bastante estudada na literatura. O cérebro é capaz de direcionar dinamicamente a atenção para locais em que características visuais relevantes são mais prováveis de estarem presentes e os estímulos que aparecem nos locais previstos são detectados mais rapidamente e com maior exatidão [Posner et al., 1980].

De maneira oposta, existe pouca investigação a respeito de orientar a atenção para um ponto particular no tempo. A Atenção Temporal se refere, portanto, ao modo como a informação sobre variações na imagem de entrada com o passar do tempo pode ser utilizada para guiar a atenção a um instante de tempo em que um evento relevante é esperado e, assim, otimizar a resposta do sistema de visão [Coull and Nobre, 1998].

Ao se tratar de Atenção Temporal, torna-se necessário referenciar um conceito importante da área de Visão Computacional: o Fluxo Óptico, que pode ser visto como um campo de velocidade posicional instantânea que associa a cada elemento do sensor visual o vetor de velocidade instantânea daquele elemento [Marr, 1982].

Fluxo Óptico

O Fluxo Óptico é definido como o movimento aparente dos padrões de brilho em uma seqüência de imagens [Yeasin, 2002]. Entretanto, essa definição, que é a mais comum, não fornece interpretações corretas em determinadas situações. Um exemplo é quando há movimento da fonte luminosa, tal como o caso que segue: um observador que se encontra parado observa uma cena também estática, iluminada por uma fonte de luz móvel. Não há movimento relativo entre o observador e a cena, mas há um Fluxo Óptico não-nulo devido ao movimento aparente do padrão da imagem.

Para se resolver tal problema, Negadharipour [Negadharipour, 1998] reviu a definição de Fluxo Óptico. Essa nova definição descreve tanto as variações radiométricas quanto as geométricas (na definição original somente as variações geométricas eram descritas). As variações geométricas se traduzem como mudanças nos

componentes X e Y do Fluxo Óptico, enquanto as radiométricas implicam diferentes intensidades (de brilho). Ou seja, na definição original, considera-se a intensidade constante, sendo o Fluxo Óptico função apenas de suas componentes horizontal e vertical. Já na nova definição, insere-se um terceiro parâmetro, a intensidade.

Existem vários algoritmos para o cálculo do Fluxo Óptico, podendo ser agrupados em três categorias: métodos baseados em gradiente, técnicas de casamento de padrões e abordagens baseadas em freqüência [Beauchemin and Barron, 1995; Barron et al., 1994]. Estudos revelam que os métodos baseados em gradiente apresentam um desempenho geral superior aos outros dois.

O Fluxo Óptico armazena informações úteis a respeito da cena. Descreve a direção e a velocidade do movimento relativo entre o observador e a cena, através de vetores (setas direcionais). As Figuras 6 e 7 apresentam um exemplo clássico de Fluxo Óptico.

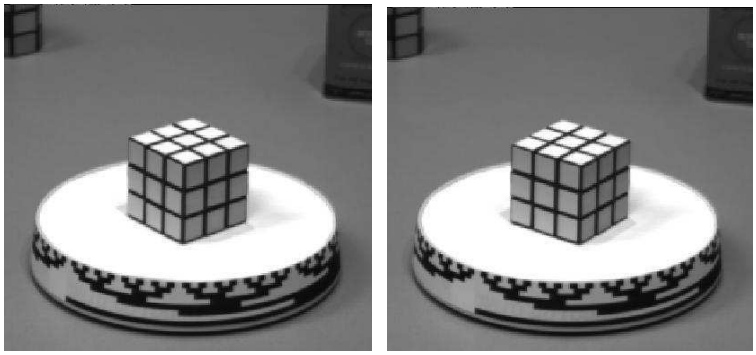


Figura 6 – Cubo de Rubik
(extraído de [Russel and Norvig, 1995])

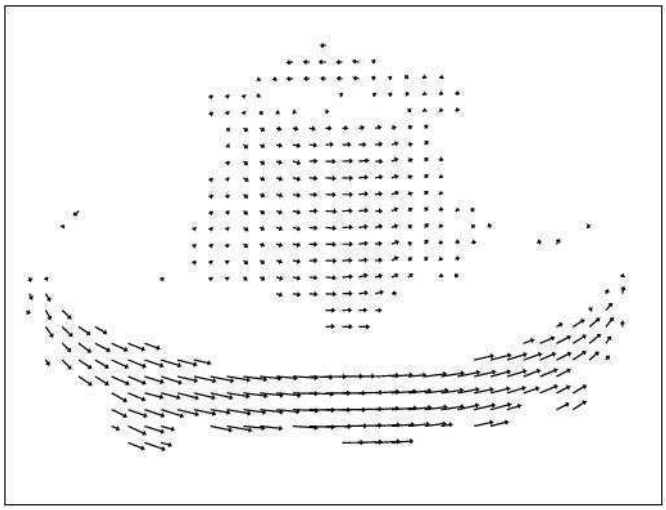


Figura 7 – Mapa de vetores de Fluxo Óptico
(extraído de [Russel and Norvig, 1995])

Na Figura 6, têm-se dois quadros de um vídeo mostrando um cubo sobre uma plataforma giratória, conhecido como *cubo de Rubik* (o primeiro quadro mostra o cubo em um momento t e o segundo mostra o cubo em um momento $t+x$). Na Figura 7, tem-se o mapa de vetores de Fluxo Óptico calculado a partir dos dois quadros da Figura 6. Os vetores indicam o movimento (módulo, direção e sentido) de um conjunto de pontos da imagem do primeiro quadro, considerando-se a posição que esses mesmos pontos estão ocupando no segundo quadro.

Classicamente, o vetor de Fluxo Óptico é representado pelas suas componentes horizontais e verticais, $v_x(x,y)$ e $v_y(x,y)$, respectivamente (consideram-se apenas as variações geométricas). Para se calcular o Fluxo Óptico, deve-se encontrar pontos correspondentes entre dois quadros consecutivos. Para tal tarefa, utiliza-se a informação de que regiões ao redor de pontos correspondentes possuem padrões de intensidade similares. O processo ocorre da seguinte maneira: Considere-se um bloco de *pixels* centrado em um *pixel p*, localizado em uma coordenada (x_0, y_0) em um tempo t_0. Esse bloco é comparado com diversos outros, centrados em *pixels* candidatos a equivalentes a p, e que estão em uma coordenada $(x_0 + D_x, y_0 + D_y)$ em um tempo $t_0 + D_t$. Nessa etapa, pode-se calcular a similaridade entre os dois blocos. Uma maneira de se fazer isso é utilizar-se da *soma dos quadrados das diferenças* (*SQD*) [Russel and Norvig, 1995]:

$$SQD(D_x, D_y) = \sum_{(x,y)} (I(x, y, t) - I(x + D_x, y + D_y, t + D_t))^2 \qquad (1)$$

em que (x, y) cobrem os *pixels* do bloco centrado em (x_0, y_0).

Encontrando-se um (D_x, D_y) que minimize a *SQD*, obtém-se o ponto correspondente (no segundo quadro), $(x_0 + D_x, y_0 + D_y)$, e o vetor de Fluxo Óptico para o ponto do primeiro quadro é definido como $(v_x, v_y) = (D_x / D_t, D_y / D_t)$.

O cálculo da disparidade para o Fluxo Óptico, acima demonstrado, pode ser aplicado também à Visão Estéreo. Entretanto, ao invés de se usar a disparidade para estimar velocidades de fluxo a partir de duas imagens adquiridas em instantes de tempo consecutivos, a disparidade em Visão Estéreo serve para estimar a profundidade da cena a partir de duas (ou mais) imagens adquiridas de pontos separados no espaço. A Visão Estéreo é discutida na próxima seção.

Conforme mencionado no capítulo introdutório desta dissertação e de acordo com a revisão bibliográfica apresentada no próximo capítulo, poucos foram os trabalhos

que exploraram a Atenção Temporal. A maioria dos trabalhos centrou-se na Atenção Espacial, principalmente a *bottom-up*. A importância de estudos que integrem evidências tanto espaciais quanto temporais para a detecção de regiões-alvo para o processo de Atenção Visual foi discutida no trabalho de Maki entre outros [Maki et al., 2000].

2.3 Visão Estéreo

Como citado na seção anterior, a idéia de Visão Estéreo é bastante semelhante à do Fluxo Óptico, só que ao invés de se usar duas imagens através do tempo, usam-se duas (ou mais) imagens separadas no espaço (provenientes de cada uma das câmeras) [Russel and Norvig, 1995]. Essas duas imagens se assemelham àquelas captadas por cada um dos olhos de um ser humano, por exemplo. Artificialmente, pode-se reproduzir tal ação usando-se duas câmeras devidamente posicionadas.

Uma dada característica de uma superfície visível estará, para cada imagem, em um local diferente, relativo ao eixo Z. Desse modo, superpondo-se as imagens, haverá uma disparidade no local dessa característica nas duas imagens. Pode-se perceber isso na figura a seguir: o ponto mais próximo da pirâmide (destacado com um círculo) está mais para a esquerda na imagem da direita e mais à direita na imagem da esquerda.

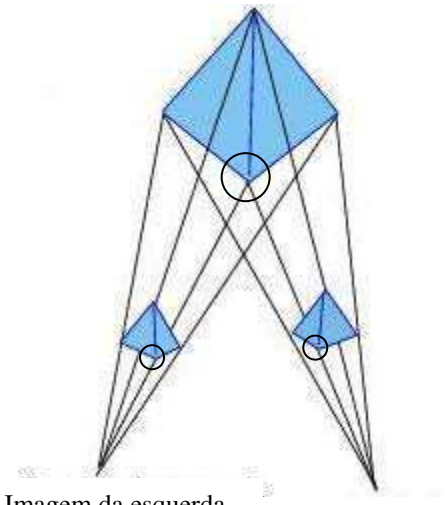


Figura 8 – Imagem original e projeções (esquerda e direita)

Ao contrário do caso do Fluxo Óptico, a Visão Estéreo necessita de algum conhecimento a respeito da geometria da visão (tipicamente, os parâmetros intrínsecos da câmera, como a distância focal, os coeficientes de distorção das lentes e a posição relativa entre as câmeras) [Russel and Norvig, 1995]. No caso do sistema visual humano, o cérebro tem conhecimento da posição dos olhos relativamente à cabeça. Do mesmo modo, em um sistema binocular de câmeras (sistema com duas câmeras), deve-se conhecer a configuração relativa.

Da mesma forma que no Fluxo Óptico, deve-se procurar pontos correspondentes nas duas imagens (cada uma recebida por uma câmera) através do cálculo da similaridade (ou da disparidade, que lhe é complementar). E esse cálculo se torna mais simples quando se possui informações a respeito da geometria (que é o que ocorre na Visão Estéreo). Pontos correspondentes devem se situar nas chamadas *linhas epipolares*, que correspondem às intersecções do plano epipolar (o plano que passa pelo ponto da cena que está focado e pelos centros das duas câmeras) com os planos das imagens da direita e da esquerda. Segue uma figura na qual são ilustradas as linhas epipolares.

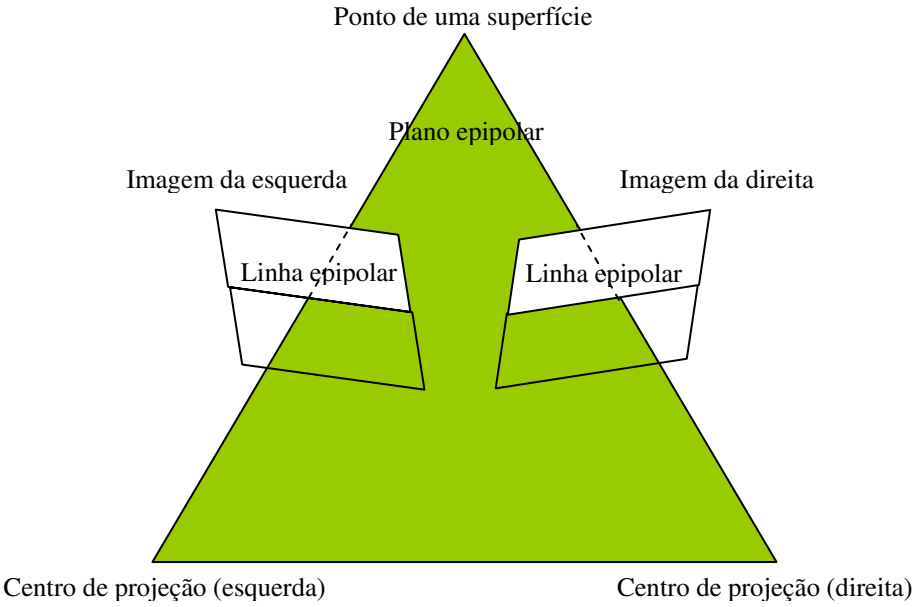


Figura 9 – Linhas epipolares

O conhecimento da localização das linhas epipolares demanda conhecimento geométrico do sistema visual e simplifica o cálculo da disparidade, que deixa de ser um

problema bidimensional e passa a ser unidimensional (a busca por pontos correspondentes fica restrita às linhas epipolares).

Tendo-se em mãos a medida da disparidade, pode-se calcular a profundidade. Observe-se a imagem apresentada na Figura 10. Considerando-se as duas câmeras voltadas para a frente e com seus eixos ópticos paralelos, tem-se que cada câmera está, em relação à outra, deslocada somente no eixo *X*, de uma distância *d*. A *linha de base* é aquela que conecta os centros das lentes das câmeras, sendo perpendicular aos eixos ópticos das câmeras e paralela ao eixo *X*. A *distância focal* (*f*), para cada câmera, é a distância perpendicular entre o centro da lente e o plano da imagem (nesse caso, as distâncias focais são iguais). O ponto *O*, localizado na linha de base e à mesma distância dos centros de cada uma das lentes, foi escolhido como origem do sistema de coordenadas. Considerando-se um ponto (x, y, z), pertencente à superfície de um objeto do mundo real, e nomeando de (x_l, y_l) e (x_r, y_r) as coordenadas das projeções desse ponto nos planos de imagem de cada uma das câmeras (esquerda e direita, respectivamente), tem-se [Russel and Norvig, 1995]:

$$x = \frac{d(x_l + x_r)}{2(x_l - x_r)} \qquad y = \frac{d(y_l + y_r)}{2(x_l - x_r)} \qquad z = \frac{df}{x_l - x_r} \qquad (2, 3, 4)$$

em que o valor da disparidade é igual a (x_l - x_r).

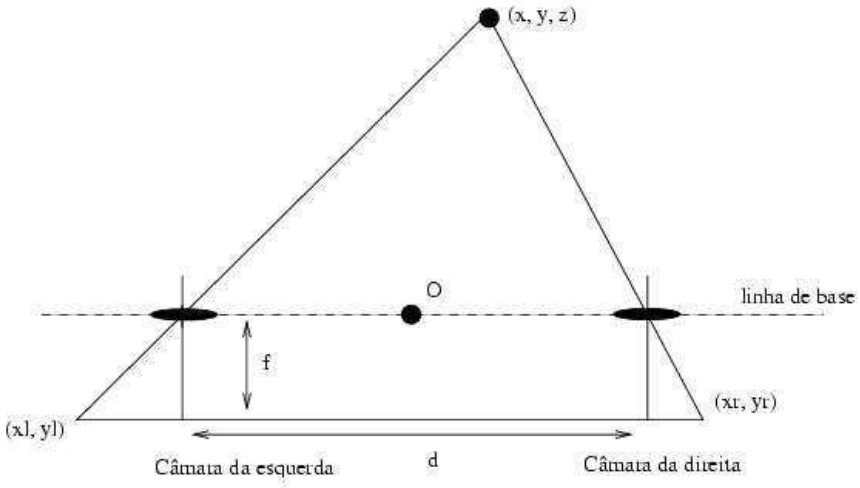


Figura 10 – Geometria de um sistema de Visão Estéreo

A forma mais simples e conveniente de representar e guardar medidas de profundidade de uma cena é através de um *mapa de profundidade*. Um mapa de

profundidade é uma imagem em tons de cinza, gerada a partir de uma cena visual da qual se consegue extrair informações sobre a profundidade, ou noção da distância relativa entre a câmera e pontos nas superfícies dos objetos. Em um sistema de Visão Estéreo de *n* câmeras, formam-se *n* imagens (cada qual associada a uma câmera diferente) e, conseqüentemente, *n* mapas de profundidade podem ser gerados. Os *pixels* de cada imagem possuem valores de profundidade associados (obtidos a partir do processo de Visão Estéreo descrito anteriormente). Em cada *pixel* do mapa de profundidade, quanto maior o valor de profundidade associado, menor a intensidade do brilho do *pixel* no mapa de profundidade. Para ilustrar, um exemplo é dado na Figura 11: à esquerda, tem-se a imagem captada pela câmera esquerda de um sistema de visão binocular e, à direita, tem-se o mapa de profundidade gerado a partir dessa imagem. Nesse mapa de profundidade, a esfera, por exemplo, está mais clara do que o cone, indicando o fato de que o cone está mais ao fundo da imagem.

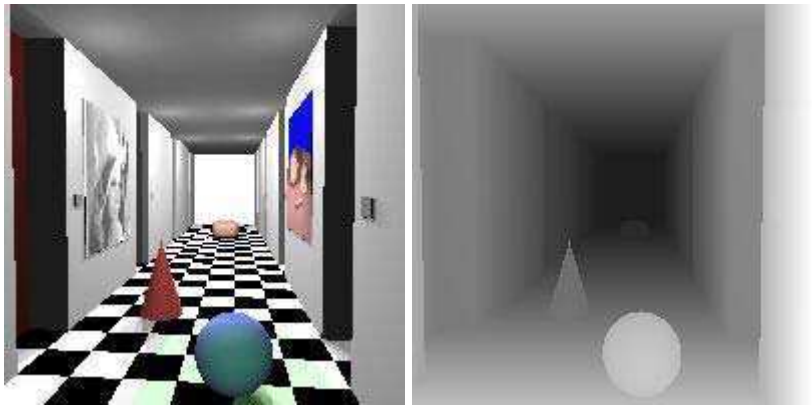


Figura 11 – Imagem da esquerda e mapa de profundidade dessa imagem (extraídas de [MRTStereo])

Capítulo 3

Trabalhos Relacionados

Neste capítulo, serão discutidos três tipos de trabalhos: trabalhos relativos aos mecanismos biológicos da Atenção Visual (Temporal ou não), que serviram como inspiração (biológica) para o desenvolvimento desta dissertação (Seção 3.1); trabalhos que tratam especificamente da Atenção Visual Espacial *bottom-up* e *top-down* (Seção 3.2); e trabalhos que lidam com cenas visuais em que há movimento (Atenção Temporal), para a execução de determinadas tarefas (Seção 3.3).

3.1 Inspiração Biológica

Uma das formas que pesquisadores em Visão Computacional têm encontrado para avançar a pesquisa em modelos computacionais de Atenção Visual (bem como em outras habilidades relacionadas à Inteligência) é estudar a estrutura e a funcionalidade de mecanismos encontrados em organismos biológicos. Já se tem um bom conhecimento sobre os mecanismos que regem a Atenção Espacial em alguns mamíferos (como gatos e macacos), a partir de estudos como os realizados por Thompson e Schall [Thompson and Schall, 2000], Pasupathy e Connor [Pasupathy and Connor, 1999], Treue e Trujillo [Treue and Trujillo, 1999], Carpenter entre outros [Carpenter et al., 1998], Roelfsema entre outros [Roelfsema et al., 1998], Logothetis entre outros [Logothetis et al., 1995], Tsotsos entre outros [Tsotsos et al., 1995], Andersen entre outros [Andersen et al., 1990], Northdurft [Northdurft, 1990], Gilbert e Wiesel [Gilbert and Wiesel, 1989], entre várias outros.

Por outro lado, ainda existem muitas lacunas com relação à Atenção Temporal. Um trabalho importante encontrado na literatura na linha de entendimento dos mecanismos biológicos da Atenção Temporal e sua relação com a Atenção Espacial foi

desenvolvido por Coull e Nobre [Coull and Nobre, 1998]. Em seu trabalho, foram identificadas quais regiões do cérebro estariam envolvidas na Atenção Temporal. Ou seja, o que se desejava saber era quais regiões do cérebro são ativadas quando se dirige a atenção a um determinado ponto no tempo, uma vez que o intervalo de tempo já foi estimado. Além disso, fez-se um estudo anatômico sobre os sistemas neurais responsáveis pela Atenção Espacial e os responsáveis pela Atenção Temporal, procurando identificar possíveis pontos em comum entre eles. Dentre as conclusões do seu trabalho, podem-se citar: (i) a confirmação da vantagem em se saber quando é mais provável que um determinado evento ocorra (e não apenas onde), (ii) a descoberta da existência de uma quantidade considerável de circuitos neurais em comum para a realização de Atenção Espacial e Temporal (apesar de haver um grau de especialização funcional em certas regiões cerebrais) e (iii) a confirmação da localização anatômica da região responsável pela Atenção Espacial (sulco intraparietal direito) e a descoberta das localizações para o caso temporal (sulco intraparietal esquerdo e córtex pré-motor inferior esquerdo).

3.2 Atenção Espacial

No trabalho de Koch e Ullman [Koch and Ullman, 1985], propôs-se um modelo de Atenção *bottom-up* que utiliza o chamado *mapa de saliência*, um mapa bidimensional que representa a saliência visual [Itti and Koch, 2001]. O mapa de saliência é utilizado pela maioria dos modelos *bottom-up* e é formado pela composição dos vários mapas de características (cada mapa de característica apresenta uma propriedade elementar da imagem, como brilho, textura, profundidade, cor, orientação etc). Tal composição gera uma medida de saliência independente de qualquer dimensão de característica. As diversas regiões da cena visual disputam nessa medida em várias escalas espaciais, e a região vencedora é eleita como a mais saliente. Dentre as dificuldades presentes na criação de um mapa de saliência, podem-se citar (i) o fato de mapas de características diferentes não serem diretamente comparáveis (escalas e natureza distintas, por exemplo) e (ii) as situações em que objetos salientes aparecem em poucos mapas de características (os objetos salientes ficarão disfarçados) ou, inversamente, objetos pouco salientes que aparecem em muitos mapas (serão realçados) [Itti et al., 1998].

Para se descobrir a região mais saliente em um mapa de saliência, pode-se utilizar uma rede neural do tipo *winner-takes-all* [Itti and Koch, 2001]. Tal rede realiza uma disputa entre os neurônios, encontrando, ao final, um único vencedor. Isso abre margem para um problema: como evitar que o mecanismo atencional aponte sempre para a mesma região (a mais saliente). Deseja-se que, após encontrada a região mais saliente, o mecanismo aponte para as demais regiões em ordem de saliência. Tal funcionalidade é obtida através de inibição de neurônios na vizinhança do neurônio vencedor, da seguinte forma: ao se encontrar a região vencedora, os neurônios de tal região são inibidos, de modo que o foco recaia sobre a segunda região mais saliente, evitando assim que pontos muito próximos sejam visitados repetidamente. O processo é repetido a cada região vencedora encontrada, fazendo com que a atenção se desloque sistematicamente para cada uma das regiões da cena, definindo assim os caminhos atencionais [Tsotsos et al., 1995].

Durante um certo tempo, havia uma maior ênfase no estudo de mecanismos de atenção do tipo *bottom-up*, ignorando-se as descobertas que estavam ocorrendo nas pesquisas de atenção baseada em objetos (*top-down*). Modelos de Atenção Visual puramente *bottom-up* não se adequam muito bem à análise de cenas complexas, ou seja, cenas onde há uma grande quantidade de objetos, que podem estar sobrepostos ou possuir propriedades em comum. Para solucionar tais casos, os modelos devem atender a três requisitos [Sun and Fisher, 2002]: funcionarem simultaneamente em regiões descontínuas; serem capazes de selecionar um objeto composto por características visuais diferentes, mas de uma mesma região do espaço; e, para objetos estruturados, poderem selecionar tanto objetos, locais e/ou características visuais, quanto seus agrupamentos.

Para se atender a esses requisitos, Sun e Fisher [Sun and Fisher, 2003] propuseram um modelo que combina Atenção *bottom-up* e *top-down*. O modelo assimila conceitos importantes, adquiridos de trabalhos de outros autores, como por exemplo: Atenção Visual baseada em objetos, segundo a Teoria da Competição Integrada de Duncan [Duncan et al., 1997; Desimone and Duncan, 1995]; saliência visual, segundo o modelo de Koch e Itti [Itti et al., 1998; Koch and Ullman, 1985]; integração das atenções *bottom-up* e *top-down* [Wolfe, 1998]; e representações visuais de "objetos dentro de objetos" (sub-componentes) e "objetos entre objetos" (sobreposição) [Humphreys, 1998]. O modelo proposto por Sun e Fisher traz como

inovações dois mecanismos: a computação de saliência baseada em agrupamentos e a seletividade hierárquica.

Com o primeiro mecanismo, a computação baseada em agrupamentos, realizam-se competições por atenção entre características, entre objetos e entre agrupamentos, além de competições dentro de objetos e dentro dos agrupamentos (de características e de objetos). As características visuais primitivas da cena são extraídas em várias resoluções, formando pirâmides de características. Calcula-se, para diferentes agrupamentos das pirâmides, a saliência visual de pontos, objetos e regiões, gerando uma competição puramente *bottom-up*. A essa competição acrescenta-se a abordagem *top-down* (comentada a seguir), obtendo-se o modelo proposto.

O segundo mecanismo, a seletividade hierárquica, serve para guiar movimentos atencionais sem movimento dos olhos (sem mudança do ponto de fixação). Analisa-se a cena em várias resoluções, inclusive os diferentes agrupamentos presentes em cada uma delas. A competição se inicia na resolução mais baixa e se desloca gradualmente para resoluções mais altas. O vencedor em cada resolução é determinado por uma estratégia do tipo *winner-takes-all*. O mesmo ocorre dentro dos agrupamentos, onde a competição por atenção move-se de agrupamentos menos refinados (a nível de pontos, ou *pixels*) para os mais refinados (objetos e regiões da cena). A conjunção dos mecanismos descritos acima produz movimentos atencionais biologicamente plausíveis, além de apresentar um desempenho de identificação de regiões importantes similar àquele produzido em experimentos com humanos.

Um problema que pode ser apontado consiste no fato da geração dos modelos (agrupamentos) de objetos ter sido realizada manualmente. Os autores realizaram tal procedimento por meio da segmentação manual dos objetos (inclusive objetos contidos em outros objetos, daí a idéia de agrupamentos) presentes nas cenas visuais. Um trabalho interessante, de forma a complementar o desenvolvido por Sun e Fisher [Sun and Fisher, 2003; Sun and Fisher, 2002], seria exatamente a geração automática dos modelos de objetos.

3.3 Atenção Temporal

Diferentemente da Atenção Espacial, que analisa imagens estáticas, a Atenção Temporal (ou Dinâmica) se preocupa em identificar o foco do processo atencional em

imagens dispostas em seqüência (vídeo). Na Atenção Espacial, tem-se uma única imagem (pode estar representada em múltiplas resoluções, mas ainda assim é uma única imagem) e o objetivo é determinar em que lugar (onde) na imagem se deve prestar atenção e qual objeto (o que) deve ser o foco da atenção. Já a Atenção Temporal objetiva saber, na seqüência de imagens, em qual delas (quando) deve recair a atenção e, nessa imagem, qual objeto (o que) merece maior destaque. Essa preocupação da Atenção Temporal com o que deve ser alvo do processo atencional denuncia as influências espaciais presentes na abordagem temporal. Ou seja, a Atenção Temporal agrega características tanto temporais quanto espaciais.

A seguir, serão revisados trabalhos que lidam diretamente com tópicos pertinentes à Atenção Temporal, como um modelo de Atenção Temporal, cálculo do Fluxo Óptico, segmentação de vídeo, detecção de movimento, renderização de vídeo, recorte automático de vídeo, detecção de transições, entre outros.

3.3.1 Modelo de Atenção Temporal

No trabalho de Ouerhani [Ouerhani, 2004], foi proposto um modelo de Atenção Temporal, seguindo os mesmos moldes do modelo de Atenção Espacial proposto por Itti [Itti, 2003; Itti, 2000; Itti and Koch, 2001; Itti et al., 1998]. Tal modelo está esquematizado na Figura 12.

O modelo se propõe a gerar um mapa de saliência de movimento para cada imagem da seqüência de vídeo. A informação de movimento utilizada por Ouerhani [Ouerhani, 2004] foi obtida através do cômputo dos vetores de Fluxo Óptico, a partir de um método baseado em gradiente multiescala. Assim, os módulos dos vetores de Fluxo Óptico (valores de velocidade) são utilizados como um indicador do movimento na cena.

Dados dois quadros consecutivos do vídeo, constrói-se uma pirâmide de resolução gaussiana para cada quadro do vídeo, através de um processo de filtragem passa-baixa e amostragem progressivas da imagem original. Ouerhani utilizou uma pirâmide em cinco níveis e calculou, para cada nível, um mapa de velocidade, obtendo, assim, a pirâmide de velocidade (as altas resoluções detectam pequenos movimentos e as baixas resoluções, grandes movimentos). Em seguida, os mapas de velocidade de cada nível são convertidos, através de um filtro passa-baixa, para a mais baixa resolução

dentre os mapas, e são somados, gerando um único mapa de saliência de movimento para o quadro de vídeo processado no momento.

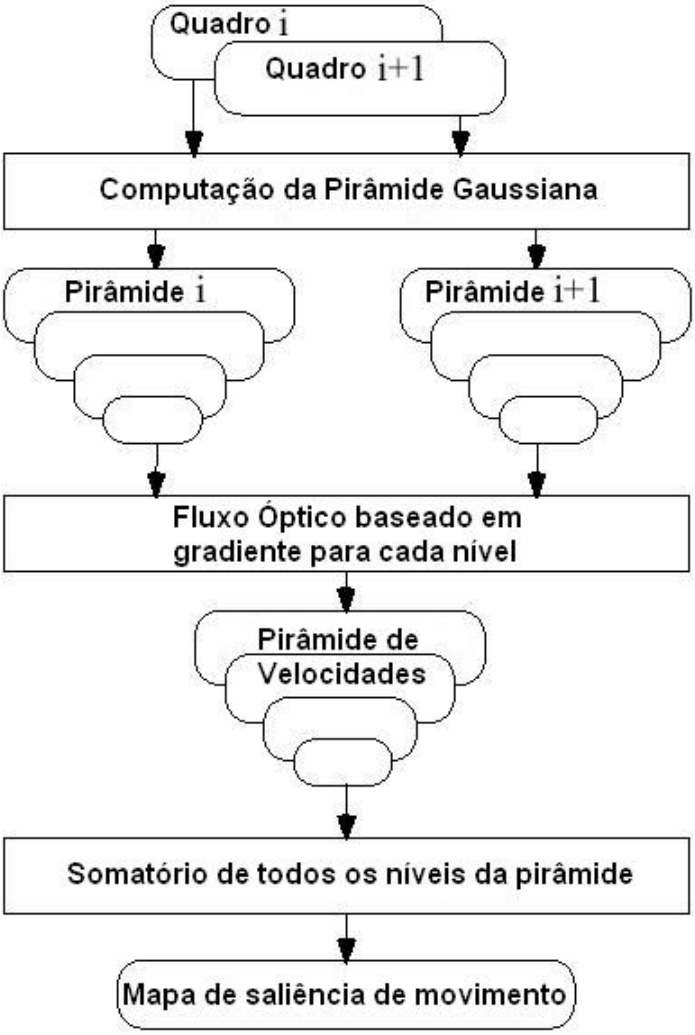


Figura 12 – Modelo de Atenção Temporal proposto por Ouerhani

Ouerhani [Ouerhani, 2004] propõe ainda a integração desse mapa de saliência de movimento com os mapas de características *bottom-up* (como intensidade, orientação e cor), gerando um mapa de saliência que integre evidências espaciais e temporais. Tal forma de integração difere da executada neste trabalho. Como será detalhado na Seção 5.4, a idéia aqui é executar um *pipeline*, ou seja, duas etapas de extração de evidências. A primeira etapa extrai as evidências temporais, ou seja, ignora as porções dos quadros do vídeo em que não ocorre movimento. Já na segunda etapa, são extraídas as características espaciais, através da geração do mapa de saliência *bottom-up* para cada

quadro do vídeo, não se levando em consideração, entretanto, todo o quadro em questão, e sim somente as regiões do quadro selecionadas na primeira etapa.

3.3.2 Cálculo do Fluxo Óptico

Dada uma seqüência de imagens (vídeo), o Fluxo Óptico consiste em um conjunto de vetores que indicam o movimento que os *pixels* de uma imagem *t* devem realizar para alcançar a posição na qual eles se encontram na imagem seguinte (*t+1*). Possui forte relação com o conceito da Atenção Temporal, por representar o movimento relativo entre o observador e a cena.

O cálculo do Fluxo Óptico se dá a partir da computação da disparidade entre dois quadros consecutivos de um vídeo (determinação dos deslocamentos relativos entre os pontos de um quadro e os pontos correspondentes no quadro seguinte). Nesse sentido, Yeasin [Yeasin, 2002] inovou com a proposição de se realizar o cálculo do Fluxo Óptico sem usar imagens convencionais (imagens cartesianas, ou seja, mapas de *pixels*). O Fluxo Óptico foi calculado a partir de imagens log-polar [Gomes and Fisher, 2003; Gomes and Fisher, 2001; Gomes et al., 1998; Rao and Ballard, 1997] geradas a partir das imagens originais. Uma imagem log-polar é uma forma de representação de imagens inspirada na representação realizada pela retina humana. A retina possui uma série de fotorreceptores distribuídos sobre uma área aproximadamente circular chamada de *campo receptivo*. A concentração dos fotorreceptores é maior na região central (fóvea), onde possuem uma alta densidade. A imagem log-polar é obtida projetando-se o ângulo e o logaritmo da distância entre o campo receptivo e o centro da retina em um sistema de coordenadas cartesianas. A seguir, descreve-se o procedimento matemático necessário para se obter essa projeção.

Considerando-se um ponto (*x*, *y*) da imagem, deve-se obter um par (ρ, θ), em que ρ é o logaritmo e θ é o ângulo. ρ e θ são obtidos a partir das fórmulas [Capurro et al., 1997]:

$$\rho = \log_b \sqrt{x^2 + y^2} \qquad \text{e} \qquad \theta = \tan^{-1} \frac{y}{x} \tag{5, 6}$$

em que $\sqrt{x^2 + y^2}$ é a distância entre a fóvea e o objeto observado e *b* controla quão rapidamente cai a densidade à medida que essa distância aumenta.

3.3.3 Segmentação de Vídeo

Dentre as técnicas aplicáveis na Atenção Temporal, pode-se citar como exemplos o Fluxo Óptico e a Detecção de Movimento. No trabalho de Maki entre outros [Maki et al., 2000], propõe-se um modelo que utiliza essas duas técnicas, juntamente com uma terceira, a Visão Estéreo (não é utilizada como técnica de Atenção Visual), para realizar segmentação da cena. Segmentação pode ser vista como o processo de organizar a matriz de *pixels* da imagem em regiões que correspondem a entidades semânticas da cena [Russel and Norvig, 1995]. O ato de segmentar e detectar movimento em cenas pode ser visto como um processo de Atenção Visual Temporal na medida em que identifica eventos importantes no tempo e pode reduzir a complexidade de processos posteriores de análise da imagem.

Uma abordagem convencional para se segmentar objetos móveis em uma cena é a subtração do fundo, utilizada, por exemplo, no trabalho de Spagnolo entre outros [Spagnolo et al., 2004]. Trata-se basicamente de se comparar um modelo de fundo com a imagem atual, detectando-se, assim, os objetos da cena que estão em primeiro plano e as formas mais confiáveis dos objetos móveis. A Equação 7 formaliza esse processo.

$$\left|I(x) - B(x)\right| > \sigma \tag{7}$$

em que $B(x)$ e $I(x)$ são os valores das imagens de fundo e atual, respectivamente, e σ é um limiar de ruído adequado.

Os algoritmos tradicionais de subtração do fundo detectam os objetos com suas próprias sombras (ou seja, não detectam a forma correta dos objetos) e geralmente usam uma abordagem baseada em *pixels* para detectar movimento. Os resultados obtidos a partir de tal tipo de abordagem não são muito bons, devido à presença de ruído e de similaridades entre o fundo e os objetos móveis [Jaraba et al., 2003].

A proposta de Spagnolo entre outros [Spagnolo et al., 2004] de um algoritmo de subtração do fundo para a segmentação de objetos móveis baseia-se na correlação existente entre *pixels* adjacentes na imagem de referência (fundo) e na imagem atual. Como as sombras dos objetos apresentam a mesma textura na imagem de referência e na atual, é possível segmentar os objetos sem suas sombras, ou seja, detectar apenas o movimento efetivo da cena (movimento dos objetos). Outros ganhos obtidos com tal

abordagem se referem ao tratamento de casos como pequenos movimentos na vegetação, mudanças graduais de iluminação (típicas de espaços abertos) e mudanças de iluminação pequenas e bruscas (por exemplo, quando uma lâmpada é acesa em um espaço fechado). O algoritmo minimiza as chances de uma região estática que está sofrendo algum desses efeitos ser erroneamente classificada como uma região móvel.

Nessa abordagem, um *pixel* é identificado como móvel comparando-se não apenas o seu valor em duas diferentes imagens (a atual e a de referência), mas também avaliando a relação de tal *pixel* com os *pixels* adjacentes. Desse modo, para se classificar um dado *pixel* como móvel ou estático, deve-se checar se sua relação com os *pixels* adjacentes se mantém substancialmente não-modificada nas imagens atual e de referência. Spagnolo entre outros [Spagnolo et al., 2004] realizaram essa checagem comparando a razão entre o *pixel* que está sendo analisado e um *pixel* adjacente na imagem atual com a razão entre os *pixels* correspondentes na imagem de referência (ver Equação 8).

$$D(i,j) = \begin{cases} \left| \dfrac{I(i,j)}{I(i,j+1)} - \dfrac{B(i,j)}{B(i,j+1)} \right| & \text{caso j < última coluna} \\ \left| \dfrac{I(i,j)}{I(i+1,j)} - \dfrac{B(i,j)}{B(i+1,j)} \right| & \text{caso j = última coluna} \end{cases} \tag{8}$$

Se $D(i, j)$ for menor que um limiar pré-definido, o *pixel* (i, j) é classificado como estático; caso contrário, como um ponto móvel. Spagnolo entre outros [Spagnolo et al., 2004] utilizaram um limiar de 0,9 (selecionado experimentalmente).

A detecção de movimento em cenas é passo fundamental no processo de segmentação de vídeo. Além disso, um módulo de detecção de movimento também foi proposto nesta dissertação (detalhado na Seção 4.2). Desse modo, comentam-se, na seção seguinte (Seção 3.3.4), alguns trabalhos que apresentaram técnicas para a detecção de movimento e que serviram de inspiração para o módulo desenvolvido neste trabalho (atenção especial dada ao trabalho desenvolvido por Wildes [Wildes, 1998]).

3.3.4 Detecção de Movimento

A detecção de objetos móveis pode ser aplicada na segmentação de vídeo (componentes móveis e fundo) e servir como foco de atenção para reconhecimento,

classificação e análise de atividades (maior eficiência, já que somente os *pixels* "móveis" são considerados), como proposto por Collins entre outros [Collins et al., 2000]. Há três abordagens convencionais para a detecção do movimento de objetos: a diferenciação temporal (bem adaptada a ambientes dinâmicos, mas ineficaz para extrair todos os *pixels* relevantes), a subtração do fundo (fornece os dados mais completos sobre os *pixels*, mas é bastante sensível a mudanças dinâmicas na cena, seja por variações na luminosidade ou por eventos externos) e o Fluxo Óptico (é capaz de detectar os movimentos de objetos e isolá-los dos movimentos da câmera, mas seus métodos são, na maioria das vezes, computacionalmente caros e complexos).

Collins entre outros [Collins et al., 2000] desenvolveram e implementaram três métodos para a detecção de movimento de objetos. O primeiro é uma combinação de subtração adaptativa de fundo e diferenciação em três quadros. Esse algoritmo híbrido é muito rápido e surpreendentemente eficaz. O segundo algoritmo foi desenvolvido para atuar em caso de falha do primeiro. Trata-se de um mecanismo para manter camadas temporais de objetos, reduzindo a ambigüidade nos casos em que objetos móveis param por um certo tempo, são encobertos por outros objetos e, então, o movimento prossegue. Uma limitação que afeta os dois primeiros algoritmos é que eles funcionam apenas com câmeras estáticas. Para superar essa limitação, propôs-se um terceiro algoritmo, que permite a subtração do fundo com câmeras móveis. Com o devido acúmulo de evidências das imagens, é possível implementar tal algoritmo em tempo real em um *PC* convencional.

A maior desvantagem da subtração adaptativa de fundo é que ela falha quando há na cena objetos estacionários que começam a se mover. Apesar de normalmente serem detectados, tais objetos deixam "buracos" nas regiões das imagens em que as novas porções do fundo expostas diferem do modelo de fundo conhecido (Figura 13a). Apesar do modelo de fundo eventualmente se adaptar a esses "buracos", eles geram alarmes falsos durante um certo período de tempo. A diferenciação de quadros não está sujeita a tal fenômeno, mas não constitui um método muito eficaz para extrair todo o formato de um objeto móvel (Figura 13b). A idéia do primeiro algoritmo proposto por Collins entre outros [Collins et al., 2000] é, então, aliar os dois métodos para suprir as deficiências de cada um. Executam-se o operador de diferenciação em três quadros, para determinar as regiões de movimento legítimo e, em seguida, a subtração adaptativa de fundo, para extrair toda a região do movimento.

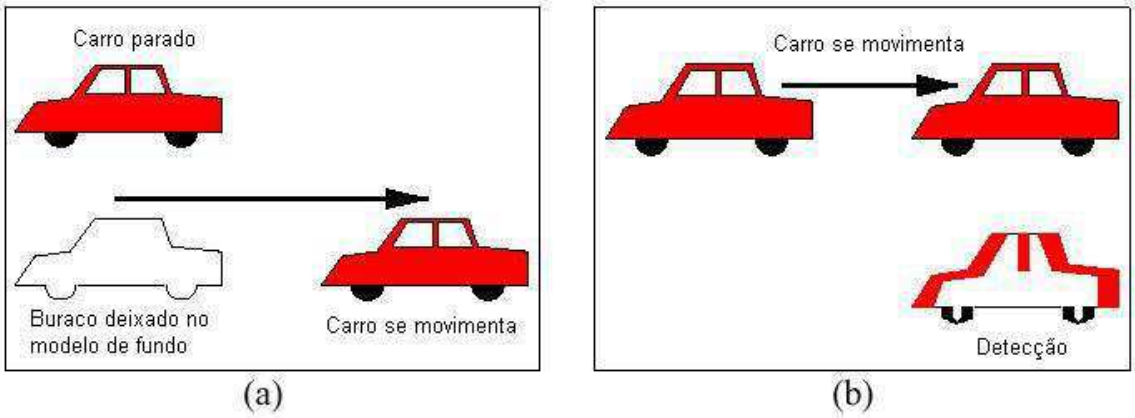


Figura 13 – Problemas apresentados pelos métodos de subtração adaptativa de fundo e de diferenciação de quadros, respectivamente

O segundo algoritmo, a detecção por camadas, é baseado na análise dos *pixels* (determina se um determinado *pixel* é estacionário ou transitório, observando sua intensidade através do tempo) e das regiões (agrupa os *pixels* em regiões estacionárias ou transitórias). Um elemento chave desse algoritmo é que ele precisa observar o comportamento de um determinado *pixel* por um certo tempo antes de determinar se ele está passando por uma transição. Observou-se que a intensidade de um *pixel* pode apresentar três perfis diferentes (ilustrados na Figura 14, em que I e t são, respectivamente, a intensidade do *pixel* e o tempo), dependendo do que está ocorrendo na cena na localização do *pixel*.

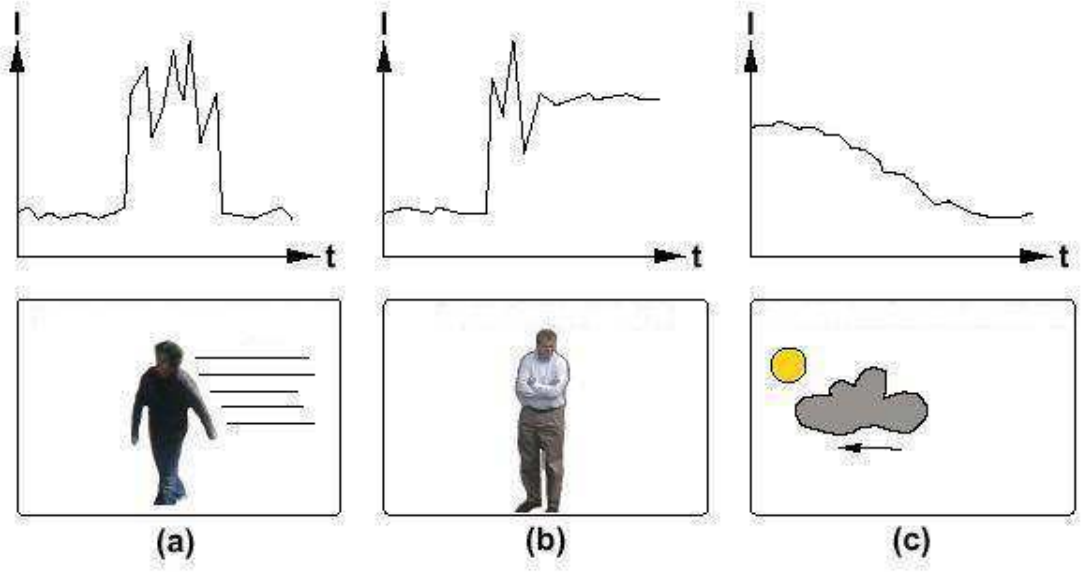


Figura 14 – Perfis apresentados pela intensidade de um *pixel*, a depender da cena

A Figura 14a apresenta uma situação em que um objeto se move e passa por um *pixel*, causando uma mudança brusca na intensidade do *pixel*, seguida por um período de instabilidade e, por fim, um retorno à intensidade original do fundo. Já na Figura 14b, um objeto se move e pára sobre um *pixel*, causando uma mudança brusca na intensidade do *pixel*, seguida por um período de instabilidade e, por fim, a intensidade se estabiliza em um novo valor, quando o objeto pára. Finalmente, na situação da Figura 14c, há a influência de efeitos meteorológicos ou luminosos, o que causa mudanças suaves na intensidade dos *pixels*.

A subtração do fundo proposta no primeiro algoritmo não é diretamente aplicável quando se está utilizando uma câmera panorâmica, já que todos os *pixels* da imagem estão se movendo. O movimento de uma câmera panorâmica é descrito, aproximadamente, como pura rotação da câmera, em que o movimento aparente dos *pixels* depende apenas do movimento da câmera e não da estrutura da cena 3D (caso bem mais simples do que se a câmera estivesse montada em um veículo se movimentando através da cena).

Collins entre outros [Collins et al., 2000] propuseram, então, um terceiro método, que utiliza um modelo de fundo completamente esférico. Há duas tarefas que precisam ser cumpridas: subtração (como a câmera se movimenta, diferentes partes do modelo esférico são recuperadas e subtraídas para revelar movimentos reais dos objetos) e atualização do fundo (à medida que a câmera visita novamente várias partes do campo de visão, as estatísticas de intensidade do fundo nessas áreas devem ser atualizadas). Ambas as tarefas dependem de se saber a direção precisa para a qual o sensor está apontando, ou seja, o mapeamento entre os *pixels* da imagem que está sendo analisada e os *pixels* correspondentes no modelo de fundo. A solução dada pelos autores para esse problema foi registrar cada imagem no modelo de fundo esférico corrente. Desse modo, tornou-se possível inferir os valores panorâmicos corretos da câmera, mesmo enquanto ela está se movendo.

Manter um modelo de fundo maior que o campo de visão físico da câmera significa representar a cena como uma coleção de imagens. No trabalho de Collins entre outros [Collins et al., 2000], um modelo de fundo inicial foi construído através da coleta metódica de um conjunto de imagens com configurações panorâmicas conhecidas. Trata-se, então, de determinar qual imagem é a apropriada, baseado na distância no

espaço panorâmico. A transformação entre a imagem corrente e uma imagem de referência aproximada consiste então em uma simples transformada de projeção planar.

O maior desafio dessa abordagem é como registrar os quadros de entrada nas imagens de referência apropriadas em tempo real. Foi desenvolvida uma nova abordagem que utiliza a integração seletiva de informação de um pequeno conjunto de *pixels* que contêm a maioria das informações sobre as variáveis de estado a serem estimadas (os parâmetros de transformação da projeção 2D). A queda dramática do número de *pixels* a serem processados resulta em uma aceleração substancial do algoritmo de registro, a ponto de ele rodar em tempo real em uma plataforma PC modesta.

Uma outra abordagem para detecção de movimento foi proposta no trabalho de Ma e Zhang [Ma and Zhang, 2001]. Em tal abordagem, a representação do movimento se baseia na imagem do espectro de energia do movimento percebido (*PMES*), obtida utilizando-se um filtro de energia temporal (elimina movimentos irrelevantes de objetos na cena) e um filtro de movimento global (diferencia os movimentos dos objetos dos movimentos da câmera). Desse modo, não se tornam necessários o uso de segmentação de objetos nem a estimação global de movimento.

Em um vídeo MPEG, existem um ou dois vetores de movimento em cada macro-bloco de um quadro P ou B, para a compensação de movimento. Esses vetores formam o campo de vetores de movimento (*MVF*). O *MVF* pode ser usado como uma aproximação do campo de Fluxo Óptico. Desse modo, está sendo privilegiada a eficiência computacional sobre a determinação exata do movimento dos objetos. Ma e Zhang [Ma and Zhang, 2001] consideraram, em seu trabalho, apenas os *MVF's* em quadros P, por questões de redução de complexidade.

O filtro de energia temporal acumula a energia ao longo do eixo do tempo. Pode-se, então, usar a magnitude dos vetores de movimento para se computar a energia do objeto ou região móvel de um macro-bloco, desde que as amostras atípicas sejam retiradas. Quanto mais intenso for o movimento de um objeto, maior será o tempo em que ele aparece em um determinado trecho do vídeo e mais fácil será de ele ser percebido por um humano. Assim, a energia de movimento em cada posição (i, j) do macro-bloco pode ser representada pela média da magnitude de movimento durante o tempo do trecho de vídeo (removendo-se amostras atípicas).

O filtro de movimento global, por sua vez, extrai a energia do movimento real dos objetos. No que diz respeito aos ângulos dos vetores de movimento, eles não representam a direção do movimento, mas a consistência espaço-temporal desses ângulos reflete a intensidade do movimento global (quanto maior a consistência, maior a intensidade). Tal consistência pode ser obtida para cada posição (*i*, *j*) do macro-bloco, medindo-se a variação do ângulo em uma janela espacial e ao longo do eixo do tempo.

Já no trabalho de Wildes [Wildes, 1998], a detecção de movimento se concretiza na forma de uma medida de saliência de movimento para aplicações de monitoramento (*surveillance*). A ênfase é dada na distinção entre movimento que tem um senso de direção coerente durante um certo espaço de tempo (como uma pessoa caminhando) e movimentos mais randômicos (como vegetação fina em uma brisa) ou periódicos (como galhos de uma árvore ao vento), ou seja, entre alvos de real interesse e meros distratores. O maior benefício dessa abordagem é que ela fornece informação sobre a saliência de movimento baseando-se em operações relativamente simples, como descrito a seguir.

O período de tempo em que movimentos randômicos ou oscilatórios se tornam aparentes é pequeno quando comparado com o período em que alvos de interesse mantêm uma direção coerente. Por isso, Wildes [Wildes, 1998] baseou a saliência de movimento na extensão na qual um determinado movimento coerente domina regiões locais no domínio espaço-temporal. Para se caracterizar tal situação, foram utilizados pares de filtros espaço-temporais que são opostos no que diz respeito à qual direção de movimento são mais sensíveis.

Considerando-se $I(x,y,t)$ como a intensidade da imagem e "*" como o operador de convolução, tem-se que as imagens:

$$R(x, y, t) = \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x} \right) * I(x, y, t) \tag{9}$$

$$L(x, y, t) = \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x} \right) * I(x, y, t) \tag{10}$$

$$U(x, y, t) = \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial y} \right) * I(x, y, t) \tag{11}$$

$$D(x, y, t) = \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial y} \right) * I(x, y, t) \tag{12}$$

fornecem medidas de movimento para, respectivamente, a direita, a esquerda, cima e baixo. A seguir, as quatros imagens são transformadas em energias espaço-temporais.

$$E_r(x,y,t) = G * [R^2(x,y,t)] \quad (13)$$

$$E_l(x,y,t) = G * [L^2(x,y,t)] \quad (14)$$

$$E_u(x,y,t) = G * [U^2(x,y,t)] \quad (15)$$

$$E_d(x,y,t) = G * [D^2(x,y,t)] \quad (16)$$

em que G é o filtro gaussiano e $A^2(x,y,t)$ é uma nova imagem formada pela potência quadrada de cada um dos *pixels* da imagem $A(x,y,t)$.

A transformação das imagens em energias atende a dois propósitos: a representação deixa de ser sensível à fase e ocorre a integração das medidas pontuais nas regiões espaço-temporais locais. A partir dessas energias, pode-se obter as medidas de saliência de movimento na horizontal e na vertical, dadas respectivamente por:

$$S_{rl}(x,y,t) = \left| \frac{E_r(x,y,t) - E_l(x,y,t)}{E_r(x,y,t) + E_l(x,y,t) + \in} \right| \quad (17)$$

$$S_{ud}(x,y,t) = \left| \frac{E_u(x,y,t) - E_d(x,y,t)}{E_u(x,y,t) + E_d(x,y,t) + \in} \right| \quad (18)$$

em que $\in$ é um número de baixo valor (Wildes [Wildes, 1998] utilizou 0,01), que serve para evitar divisão por zero em regiões sem gradiente.

Por fim, tem-se a medida global de saliência de movimento:

$$S(x,y,t) = max\,[S_{rl}(x,y,t),\, S_{ud}(x,y,t)] \quad (19)$$

3.3.5 Outros Processamentos em Vídeo

O trabalho de Yee entre outros [Yee et al., 2001] trata de dois conceitos: *sensitividade espaço-temporal* e *Atenção Visual*. A sensitividade espaço-temporal indica, para cada região de uma imagem, quanto erro se pode tolerar. Desse modo, pode-se arquitetar uma estratégia menos custosa computacionalmente para a renderização de imagens: áreas com menor sensitividade espaço-temporal (onde há

maior tolerância a erros), podem ser representadas com menor precisão (menos custo); já as áreas com maior sensitividade espaço-temporal (onde há menor tolerância a erros), seriam representadas com maior precisão (maior custo). A Atenção Visual é introduzida nesse mecanismo da seguinte forma: as áreas merecedoras de maior atenção serão representadas com maior precisão ou detalhe e, inversamente, as merecedoras de menor atenção serão representadas com menor precisão e com menos detalhes.

A iluminação global é o cálculo de quantidade de luz em um ambiente. É um cálculo computacionalmente dispendioso para ambientes estáticos e ainda mais para ambientes dinâmicos, devido à grande quantidade de imagens utilizadas e ao maior número de cálculos que devem ser realizados quando há objetos móveis. A sensitividade espaço-temporal é utilizada para se acelerar o cálculo da iluminação global, como já explicado. Entretanto, esse mecanismo não pode ser utilizado de forma ingênua: o sistema visual humano tem uma grande habilidade em rastrear objetos móveis, anulando a perda de sensibilidade devido ao movimento. É por esse motivo que se faz importante a integração da Atenção Visual nesse processo: ao se identificar as regiões da cena mais prováveis de receber atenção, identificam-se as regiões onde é mais provável de ocorrer a anulação da perda de sensibilidade.

O uso de Atenção Visual permite que tal tarefa possa ser cumprida sem o uso de mecanismos de rastreamento visual, impraticável para múltiplos observadores, mais custoso, entre outras desvantagens.

Outra aplicação para a Atenção Temporal foi proposta por Wang entre outros [Wang et al., 2004]. Em seu trabalho, foi desenvolvida uma estratégia para representação de vídeo em tempo real em pequenos aparelhos, como celulares, por exemplo. Para tanto, eles se basearam nas características espaço-temporais dos vídeos. Através da Atenção Temporal, determinaram quais eram as seqüências de interesse (quadros com maior saliência de movimento) do vídeo. Já a partir de um modelo conjunto de Atenção Espacial e Temporal, determinaram as regiões de interesse em cada um dos quadros do vídeo.

Tendo em mãos as seqüências e regiões de interesse do vídeo, Wang entre outros [Wang et al., 2004] adotaram o seguinte procedimento para a transmissão dos vídeos em pequenos aparelhos: os quadros que pertencem às seqüências de interesse são transmitidos a uma taxa de quadros igual à taxa do vídeo original. Já os quadros que não pertencem às seqüências de interesse são transmitidos a uma taxa inferior à taxa original

ou simplesmente só são transmitidos quando requisitados pelo usuário. Em seguida, cada quadro que iria ser transmitido tem sua região de maior interesse extraída e redimensionada no tamanho da tela do pequeno aparelho específico. São essas regiões de maior interesse que são transmitidas e não os quadros originais.

Outro trabalho que realizou adaptação de imagens em pequenas telas foi realizado por Chen entre outros [Chen et al., 2003]. Nele, foi proposto um modelo de Atenção Visual Espacial, segundo os moldes propostos por Itti [Itti, 2003; Itti, 2000; Itti and Koch, 2001; Itti et al., 1998]. A partir dos valores de saliência espacial, Chen entre outros [Chen et al., 2003] se propuseram a elaborar um algoritmo eficiente para adaptar imagens em pequenas telas. A idéia geral da estratégia desenvolvida no trabalho de Chen entre outros será comentada a seguir.

Primeiramente, foi definida uma métrica chamada *fidelidade de informação* (*IF*). Essa é uma métrica subjetiva, que define o quanto uma versão modificada (por compressão, recorte de imagem etc) de um determinado objeto da cena se assemelha à sua versão original. A *IF* varia entre 0 e 1 (total perda de informação e total retenção de informação, respectivamente). Assim, torna-se possível calcular a *IF* para uma determinada área *R*, contida na imagem que está sendo analisada, através da soma ponderada das *IF*'s dos objetos contidos em tal região. Segue a fórmula proposta por Chen entre outros [Chen et al., 2003].

$$IF(R) = \sum_{ROI_i \subset R} AV_i \times IF_{AO_i} \tag{20}$$

em que AO_i são os objetos contidos na região R e ROI_i e AV_i são, respectivamente, as regiões de interesse e os valores de Atenção Espacial desses objetos.

Uma vez em posse de tal métrica, o problema de adaptar uma imagem em uma pequena tela se transforma em um problema de detectar, dentro da imagem, uma região *R*, do tamanho da tela em questão, que maximize o valor da *IF* referente a essa região.

Finalmente, na área de detecção de transições, Guimarães entre outros [Guimarães et al., 2001] idealizaram um método para a detecção de transições abruptas em vídeos através do ritmo visual. O ritmo visual é uma simplificação do vídeo em uma imagem bidimensional [Guimarães et al., 2001]. Um vídeo possui três dimensões: uma temporal (correspondente aos quadros do vídeo) e duas espaciais (cada quadro é uma imagem escalar). O ritmo visual realiza uma amostragem em cada quadro do vídeo,

transformando cada quadro em uma linha vertical (uma única dimensão). É desse modo que se obtém uma representação bidimensional para os vídeos: uma temporal e outra espacial (cada quadro está em uma representação unidimensional).

Quanto à amostragem que é realizada em cada quadro para se gerar as linhas verticais, há várias possibilidades como, por exemplo, a extração da diagonal principal ou da linha vertical central de cada quadro, dentre outras. No trabalho de Chung entre outros [Chung et al., 1999], apontam-se as amostragens realizadas através da extração de diagonais como as mais interessantes, por apresentarem informações provenientes tanto das linhas quanto das colunas dos quadros.

Após a geração do ritmo visual, pode-se detectar transições abruptas no vídeo, utilizando-se técnicas de processamento de imagens bidimensionais. O procedimento consiste em identificar diferentes padrões no ritmo visual. Como cada efeito no vídeo corresponde a um padrão no ritmo visual, basta procurar pelo padrão referente a uma transição abrupta (uma linha vertical que separa dois diferentes padrões, um à esquerda e outro à direita).

Um problema nesse tipo de abordagem é que a relação 'transição abrupta - linha vertical que separa dois padrões' não é um para um. Toda transição abrupta no vídeo gera uma linha vertical que separa dois padrões no ritmo visual. Entretanto, uma linha vertical que separa dois padrões nem sempre corresponde a uma transição abrupta no vídeo, podendo representar algum outro tipo de efeito. Uma solução apresentada por Guimarães entre outros [Guimarães et al., 2001] para esse problema é realizar a busca por linhas verticais que separam padrões em diferentes ritmos visuais de um mesmo vídeo (ritmos visuais gerados a partir da utilização de diferentes amostragens do vídeo). Apenas as linhas verticais separadoras de padrões que aparecerem em todos os ritmos visuais corresponderão a transições abruptas no vídeo.

Nesta dissertação, também foi desenvolvida uma estratégia para a detecção de transições abruptas em vídeos, utilizando-se das evidências temporais obtidas a partir da geração de mapas de movimento (Seção 4.2). Essa estratégia para a detecção de transições abruptas será analisada detalhadamente na Seção 5.2.

Capítulo 4

Arquitetura do Sistema

Este capítulo apresenta uma descrição do protótipo do sistema proposto. Primeiramente, é exposta a arquitetura geral do sistema, comentando-se sucintamente sobre cada um de seus módulos e sobre como é feita a integração entre eles. Em seguida, o Módulo de Atenção Temporal é explicado mais detalhadamente, especificando-se os passos do algoritmo desenvolvido neste trabalho para a geração de mapas de movimento.

4.1 Arquitetura Geral do Sistema

O sistema modelado considera a possibilidade de *n* câmeras, que capturam, de diferentes posições, um mesmo vídeo a ser processado. O processamento de um vídeo consiste em gerar um novo vídeo em cujos quadros somente sejam visíveis as características *bottom-up* dos objetos móveis que estejam situados a, no máximo, uma distância *d* (pré-estabelecida) das câmeras. As demais regiões dos quadros são apresentadas completamente escuras (intensidade nula). Na Figura 15, pode-se visualizar uma esquematização da arquitetura proposta e, nos parágrafos que seguem, descreve-se o procedimento geral para a geração de um quadro desse novo vídeo.

Do primeiro vídeo (vídeo capturado pela primeira câmera), extraem-se o *t*-ésimo quadro (quadro *t*) e seu sucessor no tempo, o quadro *t+1*. Ambas imagens são passadas para o Módulo de Atenção Temporal, que gera o mapa de movimento. Já dos demais vídeos, são extraídos os respectivos quadros *t*. Em seguida, os quadros *t* de todos os vídeos (inclusive do primeiro) são enviados para o Módulo de Visão Estéreo, que gera o mapa de profundidade. O quadro *t* do primeiro vídeo é, então, enviado, juntamente com os mapas de movimento e de profundidade, para o Módulo de Segmentação de

Movimento e de Profundidade, que gera um quadro (Quadro Segmentado) em que apenas os objetos móveis situados a uma distância *d* (ou inferior) das câmeras são visíveis e as demais regiões são completamente escuras.

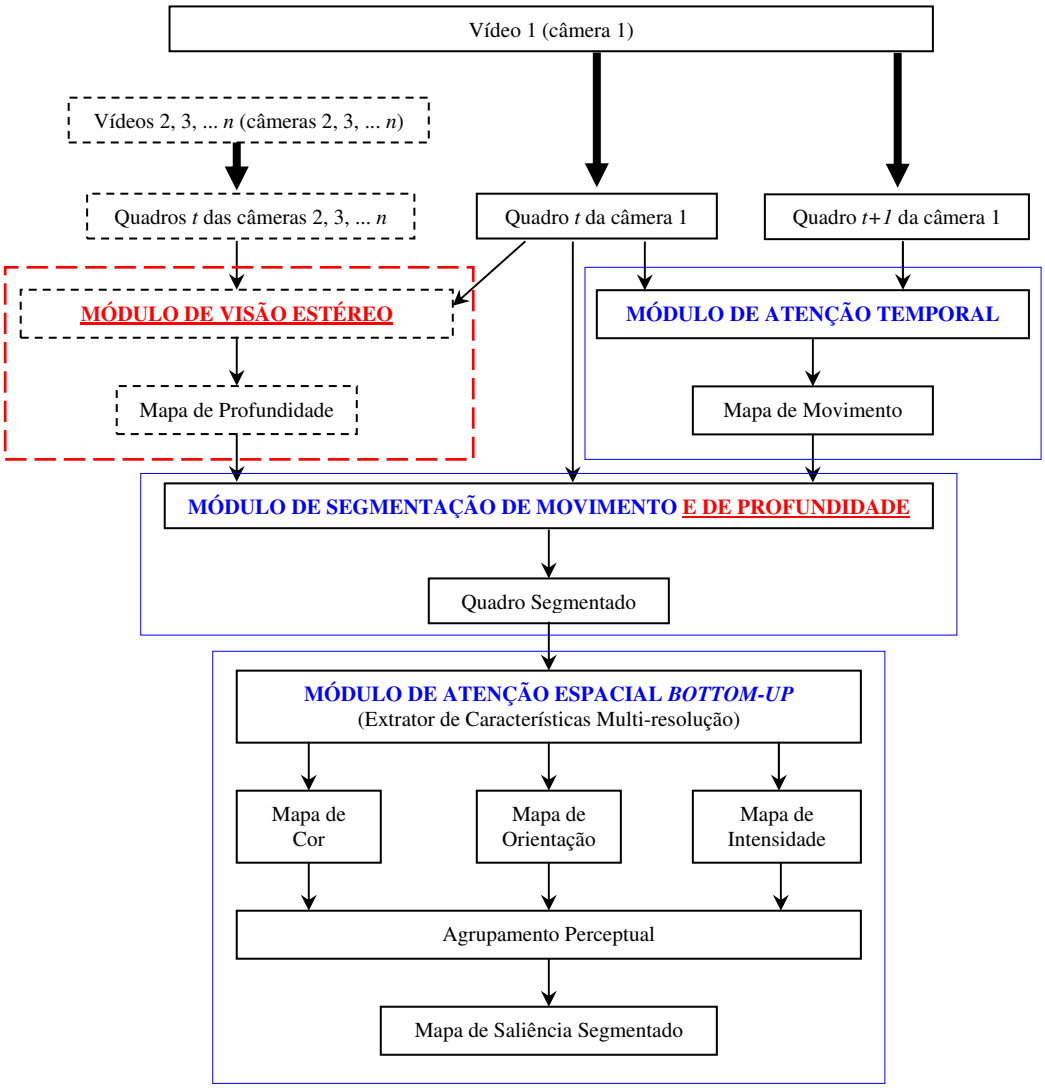


Figura 15 – Arquitetura do modelo de Atenção Visual proposto
(OBS.: Os módulos que apresentam texto sublinhado ou contorno com linha tracejada não foram implementados no sistema desenvolvido neste trabalho)

Prosseguindo com a descrição da arquitetura do sistema, o quadro segmentado pelo movimento e pela profundidade, obtido na etapa anterior, é submetido para o Módulo de Atenção Espacial *bottom-up*. Tal módulo se constitui em um Extrator de Características Multi-resolução, que recebe o quadro segmentado e o representa em várias resoluções. A seguir, são computadas as características visuais primitivas (nessa etapa, cor, intensidade e orientação) para cada resolução. Os mapas de características de

diferentes resoluções são combinados, gerando os mapas de cor, de intensidade e de orientação. Esses três mapas também são combinados, produzindo o mapa de saliência. O fator diferencial desse mapa de saliência, em comparação com os mapas produzidos em outros modelos, é que ele se constitui em um mapa de saliência *bottom-up* em que as regiões estáticas e as situadas a uma distância superior a *d* são apresentadas completamente escuras (as regiões escuras presentes no quadro segmentado pelo movimento e pela profundidade serão preservadas no mapa de saliência). Ou seja, esse mapa de saliência integra conjuntamente evidências espaciais *bottom-up*, temporais e de Visão Estéreo.

Devido a questões de escopo, o Módulo de Visão Estéreo proposto na arquitetura não foi implementado no sistema desenvolvido neste trabalho. Assim sendo, o sistema implementado conta com uma única câmera, que captura os quadros que serão enviados para o Módulo de Atenção Temporal. Já o Módulo de Segmentação se constitui apenas como um Módulo de Segmentação de Movimento, recebendo o quadro *t* do vídeo e seu respectivo mapa de movimento e gerando um quadro segmentado unicamente pelo movimento (sem segmentação pela profundidade). Por fim, o Módulo de Atenção Espacial *bottom-up* recebe esse quadro segmentado pelo movimento e gera um mapa de saliência *bottom-up* também segmentado pelo movimento.

Vale frisar que o Módulo de Segmentação de Movimento, por si só, constitui-se como uma das principais contribuições deste trabalho. Utilizando-se o sistema até a etapa de geração do quadro segmentado pelo movimento, é possível gerar, a partir do vídeo de entrada, um vídeo segmentado pelo movimento. A aparência desse vídeo é uma seqüência de imagens totalmente escuras, nas quais há alguns "clarões", como se fossem focos de luz iluminando algumas regiões de cada quadro do vídeo. Essas regiões "iluminadas" permitem o fácil rastreamento (seja humano ou automático) dos objetos móveis. Alguns resultados obtidos a partir da utilização do sistema até esta etapa e mais detalhes sobre o Módulo de Segmentação de Movimento são discutidos na Seção 5.3.

Já no que se refere ao Módulo de Atenção Espacial *bottom-up*, este segue a arquitetura geral sugerida por Itti [Itti, 2003; Itti, 2000; Itti and Koch, 2001; Itti et al., 1998] para a extração de evidências *bottom-up*. Entretanto, no caso do sistema implementado neste trabalho, esse extrator de características *bottom-up* recebe um quadro segmentado pelo movimento, ou seja, uma imagem em que apenas as regiões com algum nível de movimento estão destacadas (intensidades não-nulas). Isso agiliza a

extração das características *bottom-up* (processo pesado e lento), pois o extrator só necessita trabalhar sobre as regiões não-nulas da imagem, que constituem uma pequena minoria. Além dessa maior agilidade na extração das evidências *bottom-up*, há uma filtragem de tais evidências, pois elas só são obtidas para as regiões onde ocorre movimento. As demais regiões são completamente ignoradas (mostradas em preto) por não serem consideradas interessantes para a análise (uma vez que são regiões estáticas).

No que se refere ao aspecto implementação, a integração entre evidências espaciais e temporais proposta neste trabalho se concretiza na geração do mapa de saliência segmentado pelo movimento. Tal mapa agrega tanto características temporais, ao excluir as regiões estáticas dos quadros do vídeo ("pintando-as" de preto), quanto espaciais (extração de características *bottom-up* das regiões onde ocorre movimento).

Na próxima seção, serão apresentados maiores detalhes sobre o Módulo de Atenção Temporal do modelo proposto.

4.2 Módulo de Atenção Temporal

O Módulo de Atenção Temporal recebe como entrada dois quadros de um vídeo e gera como resposta um quadro que indica o movimento ocorrido entre um quadro e outro. Tal quadro-resposta consiste em uma imagem em tom de cinza (com a intensidade dos *pixels* entre 0 e 255), em que, quanto maior a intensidade do *pixel* (quanto mais próximo da cor branca), maior a intensidade do movimento no referido ponto da imagem; inversamente, quanto menor a intensidade do *pixel* (quanto mais próximo da cor preta), menor a intensidade do movimento. A esse quadro-resposta chama-se mapa de movimento.

O algoritmo desenvolvido e utilizado neste trabalho para a geração do mapa de movimento baseou-se naquele descrito no trabalho de Wildes [Wildes, 1998] (sumarizado na Seção 3.3.4). Segue uma descrição geral do algoritmo desenvolvido neste trabalho.

O algoritmo trabalha com uma seqüência de imagens (quadros de um vídeo) e gera uma imagem-resposta (um mapa de movimento) para cada par consecutivo da seqüência. Seja E_i, i = 1, 2, ... n, as n imagens de entrada do algoritmo, têm-se como saída as imagens $(Saída)_j$, j = 1, 2, ... (n-1), seguindo o princípio de que cada imagem

($Saída$)$_j$ é gerada a partir de cada par de imagens E_j e E_{j+1} correspondentes (esse esquema é ilustrado na Figura 16).

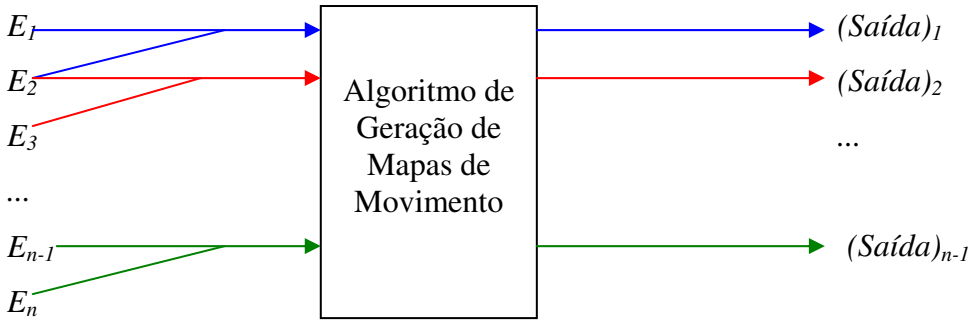


Figura 16 – Esquema geral do algoritmo de geração de mapas de movimento

Descrevem-se, a seguir, os processamentos realizados com cada par de imagens de entrada, até a geração do mapa de movimento relacionado a esse par. Essa descrição esclarecerá as adaptações deste algoritmo em relação àquele proposto por Wildes [Wildes, 1998].

Primeiramente, realiza-se uma subtração entre os dois quadros e normaliza-se o resultado em cada *pixel* para um valor entre 0 e 255. Em seguida, essa subtração normalizada sofre três convoluções, duas espaciais (uma no eixo X outra no eixo Y) e uma de atenuação.

Como o resultado da convolução de atenuação é, como o próprio nome já diz, simplesmente a atenuação da subtração normalizada (que, por si só, já representa aspectos temporais), tal convolução é tratada, no contexto deste trabalho, como uma convolução temporal. Na Figura 17, é apresentado um diagrama do processo de subtração normalizada, sendo que E_i e E_{i+1} representam cada par de quadros de entrada consecutivos e SN_i representa os resultados das subtrações normalizadas entre cada um desses pares de quadros.

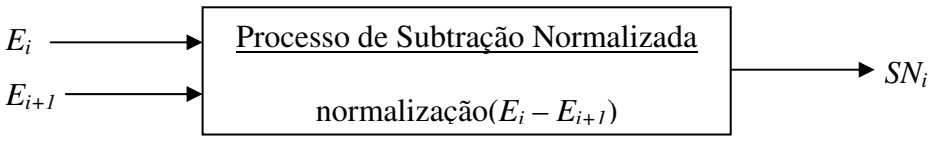


Figura 17 – Esquema geral do processo de subtração normalizada

As definições das convoluções espaciais e temporal são apresentadas nas Equações (21), (22) e (23).

$$CeX = M_x * SN \quad (21)$$

$$CeY = M_y * SN \quad (22)$$

$$Ct = M_a * SN \quad (23)$$

em que *CeX*, *CeY* e *Ct* correspondem, respectivamente, às convoluções espacial no eixo X, espacial no eixo Y e temporal, M_x, M_y e M_a são as máscaras espacial em X, espacial em Y e de atenuação e * é o símbolo de convolução.

São expostas, a seguir, as máscaras utilizadas neste trabalho para as convoluções espaciais.

$$M_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad \text{e} \quad M_y = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix} \quad (24, 25)$$

respectivamente, para os eixos X e Y.

Quanto à máscara aplicada sobre a subtração normalizada para se efetuar a convolução de atenuação (convolução temporal), tem-se:

$$M_a = \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{pmatrix} \quad (26)$$

Tendo em mãos essas três convoluções, o próximo passo é o cálculo das tendências de movimento para a direita (*R*), para a esquerda (*L*), para cima (*U*) e para baixo (*D*), transcritas da forma como foram representadas por Wildes [Wildes, 1998] para, respectivamente:

$$R(x,y,t) = Ct(x,y,t) - CeX(x,y,t) \quad (27)$$

$$L(x,y,t) = Ct(x,y,t) + CeX(x,y,t) \quad (28)$$

$$U(x,y,t) = Ct(x,y,t) - CeY(x,y,t) \quad (29)$$

$$D(x,y,t) = Ct(x,y,t) + CeY(x,y,t) \quad (30)$$

A seguir, assim como em Wildes [Wildes, 1998], essas quatros tendências (que se caracterizam como imagens ou matrizes) têm o valor de cada um de seus elementos elevado ao quadrado e sofrem uma convolução por uma função Gaussiana, sendo

transformadas nas energias espaço-temporais E_r (para a direita), E_l (para a esquerda), E_u (para cima) e E_d (para baixo).

$$E_r(x,y,t) = G * [R^2(x,y,t)] \quad (31)$$

$$E_l(x,y,t) = G * [L^2(x,y,t)] \quad (32)$$

$$E_u(x,y,t) = G * [U^2(x,y,t)] \quad (33)$$

$$E_d(x,y,t) = G * [D^2(x,y,t)] \quad (34)$$

em que G é a máscara Gaussiana e $A^2(x,y,t)$ é uma nova matriz formada pela potência quadrada de cada um dos elementos da matriz $A(x,y,t)$.

A máscara gaussiana utilizada para a computação das energias espaço-temporais é exibida na Equação 35.

$$G = \begin{pmatrix} 1/16 & 1/8 & 1/16 \\ 1/8 & 1/4 & 1/8 \\ 1/16 & 1/8 & 1/16 \end{pmatrix} \quad (35)$$

A partir das energias espaço-temporais, pode-se obter as medidas de saliência de movimento na horizontal (S_{rl}) e na vertical (S_{ud}), adaptadas das propostas por Wildes [Wildes, 1998]:

$$S_{rl}(x,y,t) = fn_1 \times \left| \frac{E_r(x,y,t) - E_l(x,y,t)}{fn_2} \right| \quad (36)$$

$$S_{ud}(x,y,t) = fn_1 \times \left| \frac{E_u(x,y,t) - E_d(x,y,t)}{fn_2} \right| \quad (37)$$

em que fn_1 e fn_2 são fatores de normalização. Neste trabalho, foram utilizados, respectivamente, os valores 4 e 510^2.

As adaptações efetuadas neste trabalho para o cálculo dessas duas medidas de saliência consistem na utilização do fator de normalização 510^2 (fn_2) como denominador das razões, enquanto Wildes [Wildes, 1998] utilizou as expressões:

$$E_r(x,y,t) + E_l(x,y,t) + \in \quad (38)$$

$$E_u(x,y,t) + E_d(x,y,t) + \in \quad (39)$$

como denominadores de S_{rl} e S_{ud}, respectivamente (só para lembrar, $\in$ é um número de baixo valor, como explicado na Seção 3.3.4), e na multiplicação de cada uma das razões por quatro (fn_1).

Essas alterações foram realizadas para se resolver problemas de geração de graves ruídos, quando da utilização das expressões como propostas por Wildes [Wildes, 1998]. Wildes efetuou a normalização das subtrações de cada par de energias espaço-temporais (direita-esquerda e cima-baixo) dividindo-os pela soma do respectivo par (o número $\in$ serve simplesmente para evitar que a soma no denominador da equação seja nula). Já neste trabalho, decidiu-se obter essa normalização através da divisão da subtração de cada par pelo valor teórico máximo que essa subtração pode assumir, ou seja, 510^2. Desse modo, os valores de S_{rl} e S_{ud} estariam, assim como os calculados por Wildes [Wildes, 1998], entre 0 e 1. Entretanto, experimentos provaram que o máximo alcançado por cada subtração em vídeos reais se situa muito abaixo do valor 510^2, o que faria com que essa normalização fornecesse valores muito baixos. Para aumentar essa faixa de valores para um patamar mais significativo, acrescentou-se, então, a multiplicação pelo fator quatro (tal valor foi determinado experimentalmente). Os valores que, após a quadruplicação, ultrapassam 1 são considerados valores destoantes em relação aos demais (que se distribuem de maneira relativamente uniforme entre 0 e 1) e são automaticamente reduzidos para 1. Em outras palavras, o processo de normalização aplicado trunca todos os movimentos de intensidade muito alta (ou muito acima da intensidade de movimento médio encontrado em vídeos reais) para o patamar 1. Em um passo mais adiante o resultado dessa normalização será novamente normalizado, dessa vez para o intervalo entre 0 e 255 (através de uma simples multiplicação por 255). Desse modo, faz-se importante essa distribuição dos valores entre 0 e 1, pois, sem tal ação, todos os valores gerados após a normalização entre 0 e 255 seriam bastante baixos, resultando, em termos de tom de cinza, em tons pretos ou muito próximos do mesmo (ou seja, os valores seriam indistinguíveis a olho nu, exceto para movimentos de amplitude extremamente elevada).

Em seguida, calcula-se a medida global de saliência de movimento (S), exatamente como proposto por Wildes [Wildes, 1998], ou seja, atribuindo a $S(x,y,t)$ o maior valor entre as saliências de movimento horizontal e vertical.

$$S(x,y,t) = max\ [S_{rl}(x,y,t),\ S_{ud}(x,y,t)] \tag{40}$$

Por fim, multiplica-se o valor de cada elemento de *S* por 255, gerando o mapa de movimento. Essa constitui uma última alteração (trivial) realizada no algoritmo em relação ao proposto por Wildes [Wildes, 1998]. O algoritmo de Wildes oferece simplesmente uma medida de saliência de movimento para cada par de *pixels* (*pixels* na mesma posição em cada uma das duas imagens de entrada). Já o algoritmo proposto neste trabalho se propõe a gerar um mapa de movimento. Como os passos anteriores garantem que os valores da matriz de saliência global se situam no intervalo de 0 a 1, tais valores podem ser facilmente normalizados entre 0 e 255, simplesmente multiplicando-os por 255. É a essa matriz de saliência global normalizada entre 0 e 255 (ou seja, em tom de cinza) que se dá o nome de mapa de movimento.

$$MM(x,y,t) = 255 \times S(x,y,t) \tag{41}$$

em que *MM* corresponde ao mapa de movimento.

A seguir, apresenta-se a descrição geral do algoritmo gerador de mapas de movimento comentado durante toda esta seção.

Algoritmo 1 – Algoritmo de geração de mapas de movimento

para $i = 1...(n-1)$ **faça** // do primeiro ao penúltimo quadro

SN = normalização($E_i - E_{i+1}$) // Soma normalizada recebe a subtração entre o quadro corrente e o quadro seguinte normalizada entre 0 e 255

$CeX = M_x * SN$ // cálculo da convolução espacial em X (convolução da soma normalizada utilizando a máscara M_x)
$CeY = M_y * SN$ // cálculo da convolução espacial em Y (convolução da soma normalizada utilizando a máscara M_y)
$Ct = M_a * SN$ // cálculo da convolução temporal (convolução da soma normalizada utilizando a máscara M_a)

// cálculo das tendências de movimento para a direita (R), para a esquerda (L), para cima (U) e para baixo (D)
$R(x,y,t) = Ct(x,y,t) - CeX(x,y,t)$
$L(x,y,t) = Ct(x,y,t) + CeX(x,y,t)$
$U(x,y,t) = Ct(x,y,t) - CeY(x,y,t)$
$D(x,y,t) = Ct(x,y,t) + CeY(x,y,t)$

// cálculo das energias espaço-temporais para a direita (E_r), para a esquerda (E_l), para cima (E_u) e para baixo (E_d): cada energia é a convolução gaussiana de uma matriz cujos elementos são as potências quadradas de cada um dos elementos da tendência de movimento específica
$E_r(x,y,t) = G * [R^2(x,y,t)]$
$E_l(x,y,t) = G * [L^2(x,y,t)]$
$E_u(x,y,t) = G * [U^2(x,y,t)]$
$E_d(x,y,t) = G * [D^2(x,y,t)]$

// cálculo das medidas de saliência de movimento na horizontal (S_{rl}) e na vertical (S_{ud})

$fn_1 = 4$

$fn_2 = 510^2$

$$S_{rl}(x,y,t) = fn_1 \times \left| \frac{E_r(x,y,t) - E_l(x,y,t)}{fn_2} \right|$$

$$S_{ud}(x,y,t) = fn_1 \times \left| \frac{E_u(x,y,t) - E_d(x,y,t)}{fn_2} \right|$$

$S(x,y,t) = max\,[S_{rl}(x,y,t),\ S_{ud}(x,y,t)]$ // a medida global de saliência de movimento corresponde ao maior valor entre as medidas de saliência de movimento na horizontal e na vertical

$MM(x,y,t) = 255 \times S(x,y,t)$ // o mapa de movimento corresponde à normalização entre 0 e 255 da medida global de saliência

$(Saída)_i = MM$ // um mapa de movimento é gerado como saída para cada par de quadros consecutivos i (quadro corrente) e (i+1) (quadro seguinte)

fim do para

Capítulo 5

Experimentos e Resultados

Neste capítulo, são descritos os experimentos realizados com os módulos do protótipo de sistema desenvolvido. Primeiramente, apresenta-se a evolução dos experimentos com o Módulo de Atenção Temporal, desde os resultados preliminares até os resultados finais. Em seguida, discute-se um estudo de caso envolvendo a detecção de transições abruptas em vídeos, utilizando-se de evidências obtidas a partir do Módulo de Atenção Temporal. Posteriormente, experimentos realizados com o Módulo de Segmentação de Movimento são apresentados e discutidos. A seguir, detalham-se os experimentos integrando Atenção Espacial (*bottom-up*) e Atenção Temporal. Por fim, são fornecidos detalhes de implementação. Vale frisar que a otimização do desempenho computacional dos algoritmos e estratégias propostos não constitui um objetivo final deste trabalho e que os experimentos realizados não ocorrem em tempo real.

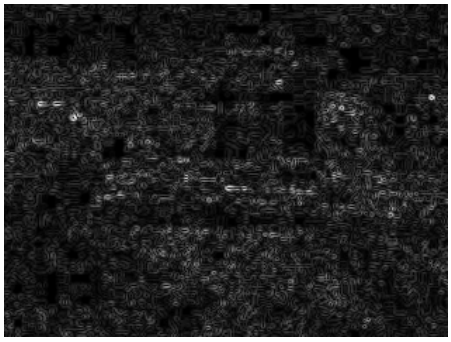
5.1 Atenção Temporal

Experimentos para o cálculo de Atenção Temporal foram realizados a partir da geração de mapas de movimento, utilizando-se o algoritmo descrito na Seção 4.2. Primeiramente, foram capturados vídeos a partir de uma *webcam*. Os quadros desse vídeo foram extraídos utilizando-se o *software* Adobe Premiere [Adobe Premiere] e passados para o algoritmo, que gera um mapa de movimento para cada par consecutivo de quadros.

Primeiramente, os mapas de movimento apresentavam muito ruído e percebia-se que, quanto menor a quantidade de movimento na cena, pior era o ruído. As Figuras 18 e 19 apresentam exemplos de cenas estáticas ou praticamente estáticas e o grau de ruído

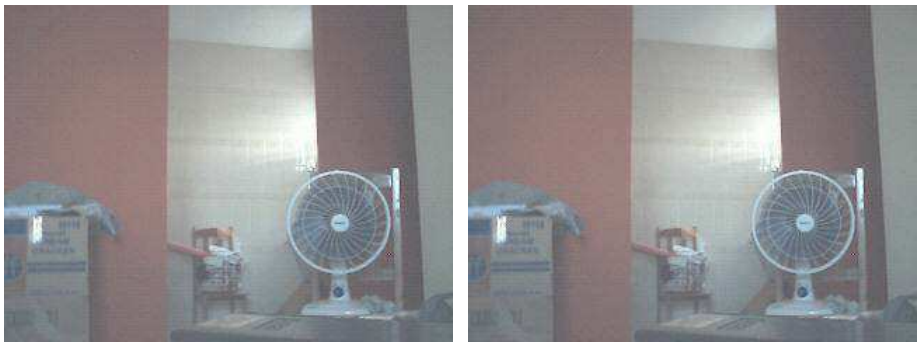
ocasionado por tal situação. Cada exemplo consiste em três imagens, as duas primeiras sendo o par de quadros e a terceira, o mapa de movimento correspondente.

(a)

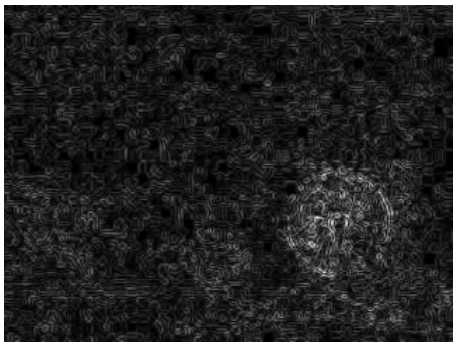


(b)

Figura 18 – Exemplo 1 de Atenção Temporal: (a) par de imagens de entrada; (b) imagem de saída



(a)



(b)

Figura 19 – Exemplo 2 de Atenção Temporal: (a) par de imagens de entrada; (b) imagem de saída

A geração desse ruído foi solucionada a partir de mudanças no algoritmo originalmente proposto por Wildes [Wildes, 1998], mais especificamente no que diz respeito à parte de normalização de valores (Seção 4.2). O fato é que as primeiras tentativas de normalização faziam com que os *pixels* em movimento, inclusive os *pixels* referentes a ruídos, apresentassem altas intensidades (tanto os *pixels* em movimento quanto os ruídos ficavam bem visíveis). Após alguns ajustes nesse mecanismo de normalização, os ruídos passaram a apresentar baixas intensidades (bem próximas a zero), ou seja, passaram a apresentar, em escala de tom de cinza, intensidades escuras, indistinguíveis a olho nu da intensidade zero (referente aos *pixels* estáticos). Já os *pixels* em movimento apresentaram intensidades médias ou altas, a depender de quão intenso fosse o movimento do *pixel*.

Um outro problema detectado pelos experimentos foi a geração, em intervalos regulares dentro da seqüência de mapas de movimento gerados, de mapas de saliência completamente ou quase que completamente formados por intensidades nulas, até mesmo em seqüências de vídeo que não apresentavam nenhum quadro sem objetos móveis. Um exemplo desse problema é mostrado na Figura 20.

(a)

(b)

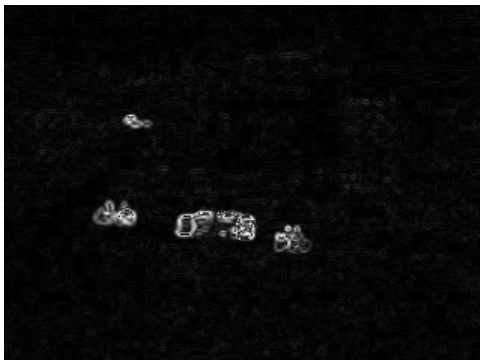
Figura 20 – Exemplo 3 de Atenção Temporal: (a) par de imagens de entrada; (b) imagem de saída

Constatou-se que tal problema acontecia quando o vídeo era capturado a uma alta taxa (por exemplo, 24 quadros por segundo). O problema foi resolvido capturando-

se o vídeo a uma taxa menor (seis quadros por segundo, por exemplo) ou usando-se mesmo um vídeo a 24 quadros por segundo, mas utilizando-se quadros em intervalos de cinco em cinco quadros (por exemplo) para a geração dos mapas de movimento. Acredita-se que a geração de quadros adjacentes idênticos durante o processo de captura de um vídeo seja um artefato do filtro de compressão utilizado, no caso deste trabalho, o filtro MPEG 1. Quando não há mudanças substanciais de um quadro para outro, é provável que o algoritmo de compressão considere que o par de quadros é idêntico. Quando a taxa de captura é reduzida, as mudanças entre quadros consecutivos tende a aumentar, minimizando a chance do problema ocorrer.

Uma vez que os dois problemas citados foram solucionados, chegou-se ao resultado desejado: mapas de movimento (i) com uma boa representatividade do movimento existente na cena (quanto maior a intensidade do movimento, mais próxima do branco é a intensidade do *pixel*), (ii) com ruídos suavizados e (iii) que só se apresentam completamente pretos em cenas verdadeiramente estáticas. As Figuras 21 e 22 ilustram os melhoramentos conseguidos.

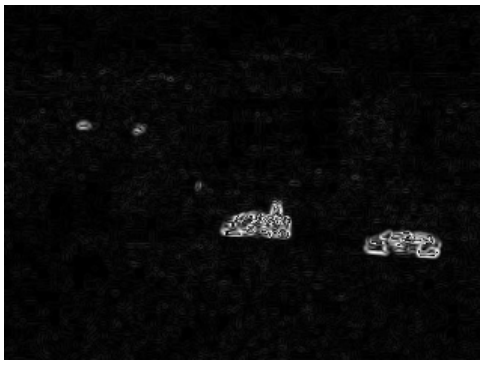
(a)



(b)

Figura 21 - Exemplo 4 de Atenção Temporal: (a) par de imagens de entrada; (b) imagem de saída

(a)



(b)

Figura 22 – Exemplo 5 de Atenção Temporal: (a) par de imagens de entrada; (b) imagem de saída

Na Figura 21, percebe-se claramente o movimento de três veículos na parte anterior da imagem (da esquerda para a direita: uma motocicleta, um automóvel e uma bicicleta) e um quarto veículo (não identificável) no *background*. Já na Figura 22, percebe-se claramente os dois automóveis em movimento, um pedestre próximo ao automóvel mais à esquerda, além de dois outros veículos se movimentando numa rua ao fundo (mais distante da câmera).

5.2 Aplicação de Atenção Temporal na Detecção de Transições em Vídeo

Nesta seção, apresenta-se um estudo de caso que demonstra a aplicabilidade do mecanismo de Atenção Temporal desenvolvido na detecção automática de transições abruptas em vídeos genéricos. A detecção de transições é um processo importante para o processamento de vídeo, especialmente para a segmentação no tempo. Pode haver dois tipos de transições em um vídeo: as graduais e as abruptas [Guimarães et al., 2001]. Neste trabalho, foca-se na detecção de transições abruptas.

A idéia é que, através da Atenção Temporal, sejam geradas medidas de movimento global para os quadros do vídeo. Ao se analisar e quantificar as transições entre essas medidas de movimento, torna-se possível detectar transições abruptas no vídeo, seguindo um preceito simples: grandes transições entre as medidas de movimento globais correspondem a transições abruptas no vídeo.

Uma pequena modificação no algoritmo gerador de mapas de movimento, comentado na Seção 4.2, executa uma soma dos valores das intensidades dos *pixels* de cada um dos mapas de movimento (imagens em tom de cinza, cada uma se referindo a um par consecutivo de quadros do vídeo). Como a intensidade de um *pixel* em um mapa de movimento representa a saliência de movimento naquele *pixel*, essa soma fornece, para cada um dos mapas de movimento, uma medida de movimento global.

De posse das medidas de movimento global para um determinado vídeo, pode-se desenhar o gráfico do comportamento do movimento durante o vídeo (Figura 23). Desse modo, a derivada de primeira ordem de tal gráfico (Figura 24) representa as variações do movimento (variações das medidas de movimento global), enquanto a derivada segunda exibe as amplitudes das variações do movimento durante o vídeo. A estratégia proposta neste trabalho para a detecção das transições abruptas consiste na análise dos módulos dessas amplitudes (módulos das derivadas de segunda ordem, ver Figura 25). Valores altos de amplitude em módulo seguidos e/ou antecedidos por valores baixos (ou seja, picos no gráfico do módulo da derivada segunda) indicam transições abruptas.

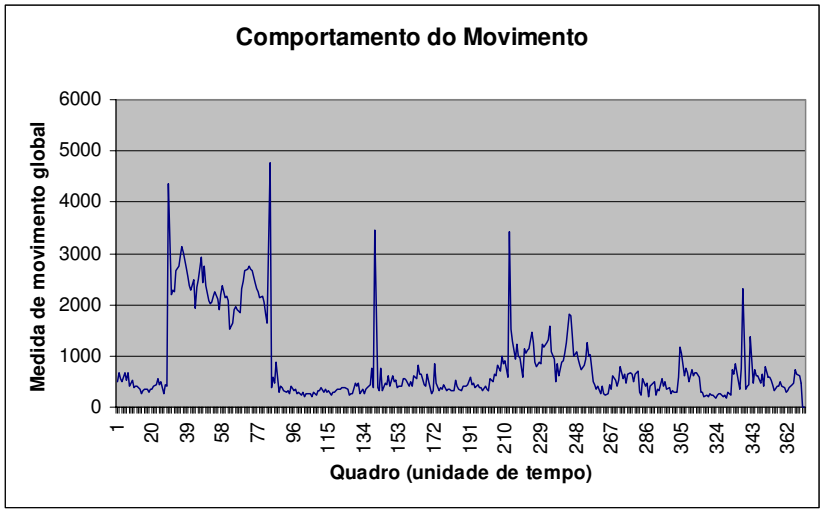


Figura 23 – Gráfico do comportamento do movimento existente em um vídeo usado como exemplo

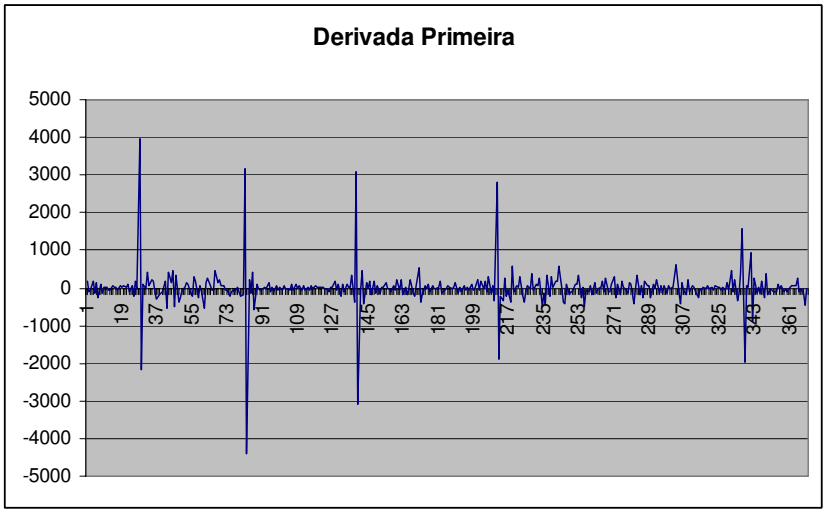


Figura 24 – Derivada primeira do gráfico da Figura 23

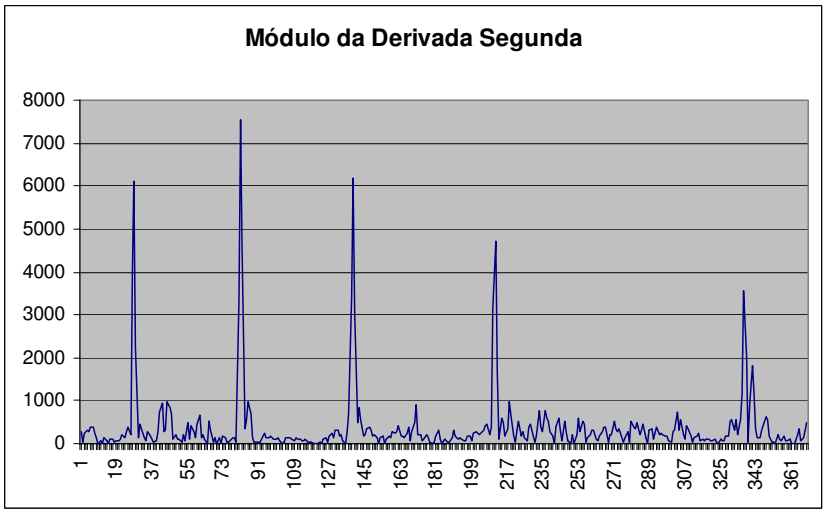


Figura 25 – Módulo da derivada segunda do gráfico da Figura 23

Foi definida a seguinte estratégia geral para se detectar transições abruptas: (i) comparar cada valor do módulo da derivada segunda com os valores ao seu redor; (ii) se o valor em questão for alto e os valores anteriores ou os valores posteriores forem baixos, os quadros do vídeo referentes ao valor em questão são considerados transições abruptas. Cabe agora especificar os detalhes de como são feitas tais comparações: cada valor é comparado com a média dos *N* valores a partir do primeiro vale (ponto de mínimo local) à esquerda e com a média dos *N* valores a partir do primeiro vale à direita. Se o valor em questão for maior ou igual a *K* vezes a média da esquerda ou a *K* vezes a média da direita, os quadros referentes ao valor em questão são considerados transições abruptas.

Basta agora definir quais os valores ideais para *N* e *K*. Para tanto, foi realizado o seguinte experimento: para um conjunto de cinco vídeos (conjunto de treinamento), foi realizada, primeiramente, uma detecção manual das transições abruptas (foram anotados, para cada vídeo, os quadros considerados como transições abruptas). No Apêndice A, Figuras 46 a 50, pode-se encontrar uma visualização dos quadros-chave de cada um desses vídeos. Os vídeos se constituem em seqüências capturadas, respectivamente, (i) de uma série de ação (apesar de, no vídeo capturado, praticamente só aparecerem pessoas caminhando e conversando), apresentada na Figura 46; (ii) de um programa cômico, no qual ocorre muito movimento (Figura 47); (iii) uma série de entrevistas com diferentes pessoas em diferentes locais (normalmente as pessoas permanecem paradas, apenas conversando; as transições ocorrem quando a cena muda de uma pessoa entrevistada para outra), apresentada na Figura 48; (iv) de uma partida de basquete (o vídeo capturado corresponde ao momento de um arremesso livre, então os jogadores permanecem a maior parte do tempo sem realizar movimentos intensos; as transições só ocorrem quando muda a câmera que está transmitindo o lance), apresentada na Figura 49, e (v) de um comercial de carro, em que o carro passa em diferentes cenas e, em cada uma delas, as pessoas estão praticando diferentes ações (Figura 50).

A seguir, foi executada a detecção automática das transições abruptas nos cinco vídeos, conforme estratégia discutida ao longo desta seção. Foram utilizados o valor de *N* variando de 1 a 20, com a variação de uma unidade, e o valor de *K* variando de 1,0 a 80,0, com a variação de 0,5. Para cada um dos vídeos e cada uma das combinações dos valores de *N* e *K*, foram calculadas as seguintes métricas (em percentual):

$$RFR = \frac{n_{FR}}{N_M} \times 100 \quad \text{e} \quad RFA = \frac{n_{FA}}{N_M} \times 100 \qquad (42, 43)$$

em que *RFR* e *RFA* significam, respectivamente, *Razão das Falsas Rejeições* e *Razão das Falsas Afirmações*, e n_{FR}, n_{FA} e N_M são o número de falsas rejeições (número de quadros do conjunto selecionado manualmente que não foram detectados pelo algoritmo), o número de falsas afirmações (número de quadros apontados como transições pelo algoritmo, mas que não constam no conjunto selecionado manualmente) e o número total de quadros do conjunto selecionado manualmente, respectivamente.

Inicialmente, foram utilizados valores de *K* variando no mesmo intervalo utilizado para os valores de *N* (1 a 20). Esse intervalo foi sendo gradativamente aumentado durante os experimentos, no intuito de alcançar o pior resultado médio possível para a *RFR* (ou seja, a taxa média de 100%). Uma taxa bem próxima de 100% foi encontrada quando da utilização do intervalo de 1,0 a 80,0 para os valores de *K*, sendo então este o intervalo escolhido para a execução dos experimentos. O objetivo em se atingir a pior taxa média foi fornecer um maior poder de representação aos resultados obtidos a partir dos experimentos, através da obtenção não apenas da melhor taxa possível (o procedimento para a obtenção da melhor taxa média para a *RFR* – e também da melhor taxa média para a *RFA* – será detalhadamente descrito um pouco mais adiante, ainda nesta seção), mas também da pior (100%).

Vale notar que a *RFA*, ao contrário da *RFR*, não apresenta um intervalo de valores entre 0 e 100%. Isso ocorre porque o número de falsas afirmações geradas pelo sistema pode ultrapassar o número total de quadros do conjunto selecionado manualmente. Tal caso não ocorre no caso da *RFR*, pois todas as falsas rejeições estão contidas no conjunto selecionado manualmente. Foi por causa desse fato que não houve uma preocupação, neste trabalho, em se encontrar intervalos de variação para os valores de *N* e *K* de modo a se obter uma taxa média de 100% para a *RFA*, (como foi realizado para a *RFR*, explicado anteriormente), já que a *RFA* pode atingir valores superiores a 100% e não apresenta um valor máximo fixo (o valor máximo teórico da *RFA* depende do número de quadros do vídeo e do número de transições abruptas detectadas manualmente).

Para maior facilidade de manuseio, cada combinação dos valores de *N* e *K* foi representada como um número inteiro *i*, que varia de 1 até *n* (número de combinações entre os valores de *N* e *K*, ou seja, $n = 3180$). Essa equivalência e a fórmula geral para a obtenção de um valor de *i* a partir de um par de valores *N* e *K* são apresentadas, respectivamente, nas Equações 44 e 45.

$$
\begin{aligned}
&i = 1 \quad \Rightarrow \quad K = 1{,}0 \text{ e } N = 1 \\
&i = 2 \quad \Rightarrow \quad K = 1{,}0 \text{ e } N = 2 \\
&i = 3 \quad \Rightarrow \quad K = 1{,}0 \text{ e } N = 3 \\
&\ldots \\
&i = 20 \quad \Rightarrow \quad K = 1{,}0 \text{ e } N = 20 \\
&i = 21 \quad \Rightarrow \quad K = 1{,}5 \text{ e } N = 1 \\
&i = 22 \quad \Rightarrow \quad K = 1{,}5 \text{ e } N = 2 \\
&\ldots
\end{aligned}
\tag{44}
$$

...
$i = 3178 \Rightarrow K = 80,0 \text{ e } N = 18$
$i = 3179 \Rightarrow K = 80,0 \text{ e } N = 19$
$i = 3180 \Rightarrow K = 80,0 \text{ e } N = 20$

$$i = 40 \times (K-1) + N \quad (45)$$

Desse modo, foram gerados, para cada um dos vídeos, dois conjuntos:

$$CFR_v = \{RFR_{1,v}, RFR_{2,v}, ..., RFR_{n,v}\} \quad (46)$$

$$CFA_v = \{RFA_{1,v}, RFA_{2,v}, ..., RFA_{n,v}\} \quad (47)$$

em que CFR_v e CFA_v significam, respectivamente, *Conjunto das Falsas Rejeições* e *Conjunto das Falsas Afirmações* para um dado vídeo v, e $RFR_{i,v}$ e $RFA_{i,v}$ são, respectivamente, os valores da razão das falsas rejeições e da razão das falsas aceitações para uma dada combinação de N e K (representada por i) e vídeo v.

Em seguida, foi realizada a média aritmética das $RFR_{i,v}$ e das $RFA_{i,v}$ (i variando de 1 até n) sobre os cinco vídeos:

$$MFR_i = \frac{1}{5}\sum_{v=1}^{5} RFR_{i,v} \quad (48)$$

$$MFA_i = \frac{1}{5}\sum_{v=1}^{5} RFA_{i,v} \quad (49)$$

em que MFR_i e MFA_i são, respectivamente, as médias das razões das falsas rejeições e das falsas afirmações para um dado i (uma dada combinação de um valor de N com um valor de K).

Desse modo, dois novos conjuntos foram obtidos, conforme as Equações 50 e 51.

$$CMFR = \{MFR_1, MFR_2, ..., MFR_n\} \quad (50)$$

$$CMFA = \{MFA_1, MFA_2, ..., MFA_n\} \quad (51)$$

em que *CMFR* e *CMFA* significam, respectivamente, *Conjunto das Médias das Falsas Rejeições* e *Conjunto das Médias das Falsas Afirmações*.

De posse desses dois conjuntos, o próximo passo foi analisar que valor de i (ou seja, que dupla de valores N e K) gera os menores valores de *MFR* e *MFA*, isto é, que

valor de *i* gera, em média, o menor número de falsas rejeições e de falsas afirmações. Uma vez encontrado esse valor de *i*, ele seria utilizado como parâmetro do algoritmo desenvolvido neste trabalho para se detectar transições abruptas em outros vídeos.

Para observar a variação das MFR_i e MFA_i, foram desenhados os gráficos *i x MFR* e *i x MFA*, expostos nas Figuras 26 e 27. As Figuras 28 e 29 mostram ampliações (ou *zooms*) de pequenas faixas, respectivamente, das Figuras 26 e 27, para melhor visualização da tendência de cada um dos gráficos. O intervalo escolhido para se efetuar o *zoom*, em ambos os gráficos, foi de *i* variando de 320 a 390.

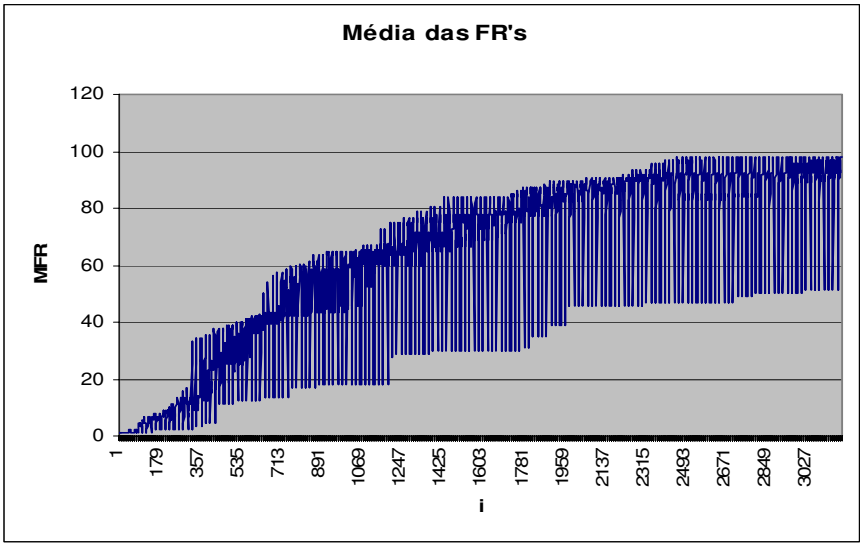


Figura 26 – Gráfico i x MFR

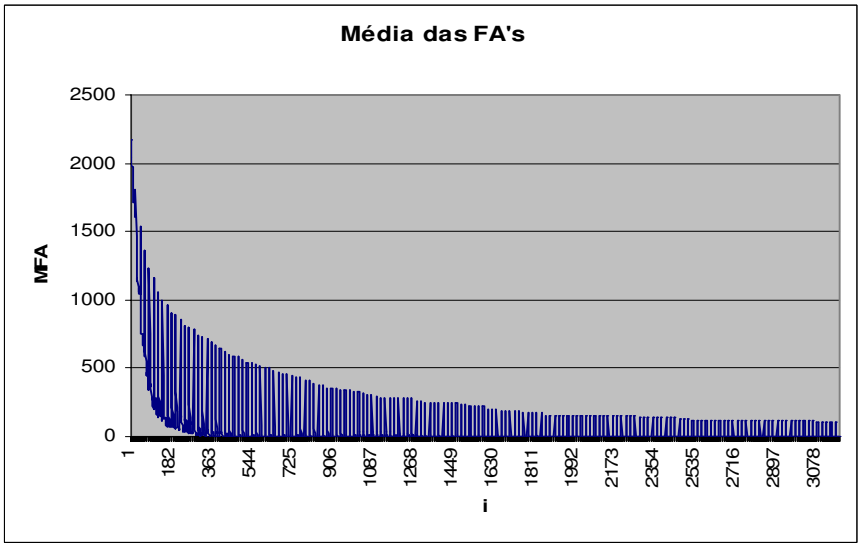


Figura 27 – Gráfico i x MFA

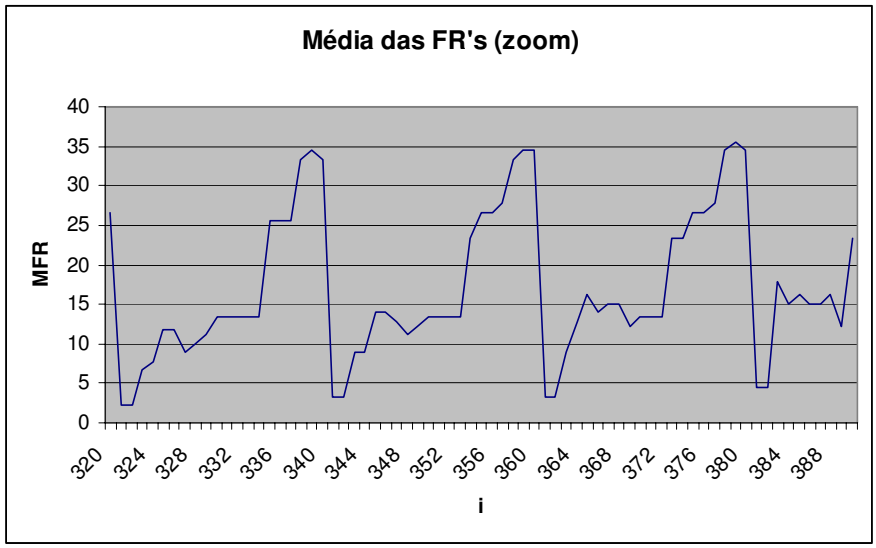


Figura 28 – Zoom do gráfico da Figura 26

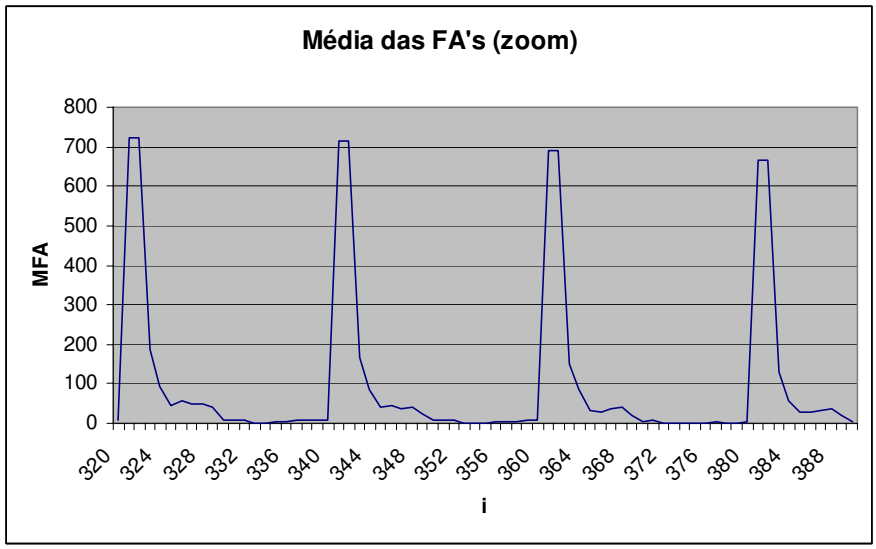


Figura 29 – Zoom do gráfico da Figura 27

Analisando-se os gráficos das Figuras 28 e 29, pode-se perceber as oscilações dos valores das MFR_i e MFA_i. Tais oscilações se explicam pela característica do valor *i*: no intervalo de 1 a 20 para *i*, o valor de *K* é 1,0 (constante) e o valor de *N* varia também de 1 a 20; já *i* assumindo valores de 21 a 40, o valor de *K* é 1,5 (constante) e o valor de *N* varia novamente de 1 a 20; e assim por diante. O movimento oscilatório é explicado a seguir. Dado um determinado valor constante de *K*, à medida que o *N* vai aumentando (de 1 a 20), o valor de MFR_i tende a aumentar e o valor de MFA_i tende a diminuir. Isso ocorre porque quanto maior o *N* (número de quadros que serão utilizados na média), mais rigorosa se torna a detecção, ou seja, mais elevada deve ser a intensidade de

movimento do quadro que está sendo avaliado, para que este seja apontado como uma transição abrupta pelo sistema, fato este que favorece a ocorrência de mais falsas rejeições e menos falsas afirmações. Quando K sofre um aumento de 0,5, em relação a seu valor anterior, N volta a apresentar o valor 1, o que faz com que o valor de MFR_i, que estava aumentando, sofra uma brusca diminuição, e o valor de MFA_i, que estava diminuindo, sofra um brusco aumento. Com o passar dos valores de N de 1 a 20, os valores de MFR_i e MFA_i tendem novamente a, respectivamente, aumentar e diminuir.

Já pelos gráficos das Figuras 26 e 27, percebe-se claramente que, apesar do comportamento oscilatório, os gráficos $i \times MFR$ e $i \times MFA$ apresentam tendências crescente e decrescente, respectivamente. Ou seja, cada brusca(o) diminuição (aumento) que o valor de MFR_i (MFA_i) sofre, após um ciclo de variações do valor de N, tende a gerar um valor superior (inferior) ao valor apresentado após o ciclo anterior. Desse modo, os menores valores de *MFR* são dados por baixos valores de i, enquanto que os menores valores de *MFA* são dados por altos valores de i. Assim, faz-se necessário encontrar um valor intermediário de i que equilibre os valores de *MFR* e *MFA*. Para tal, foi calculada a média ponderada dos valores de *MFR* e *MFA*, como segue.

$$MP_i = (w_1 \times MFR_i) + (w_2 \times MFA_i) \tag{52}$$

em que MP_i é a média ponderada dos valores de MFR_i e MFA_i e $w_1 + w_2 = 1{,}0$ são os pesos associados a cada medida de erro.

A escolha dos pesos (w_1 e w_2) deve variar dependendo da aplicação (prioridade em reduzir as falsas rejeições ou as falsas afirmações, em particular). Para os experimentos realizados neste trabalho, optou-se por se utilizar pesos iguais para essas duas métricas ($w_1 = w_2 = 0{,}5$), não favorecendo a redução de uma única métrica, em detrimento da outra. Ou seja, foi executada a média aritmética simples dos valores de *MFR* e *MFA*. A Figura 30 apresenta o gráfico $i \times MP$, utilizando tais pesos, e a Figura 31, um *zoom* desse gráfico, no intervalo de i variando de 320 a 390.

O valor de i do gráfico da Figura 30 que corresponde ao menor valor de *MP* é o resultado que tem sido procurado até o momento, ou seja, um valor de i interessante para se utilizar na detecção de transições abruptas em vídeo. No caso, há três valores diferentes de i que fornecem a menor *MP* do gráfico (destacados na Figura 31): 334, 353 e 372. Tais valores resultam nos seguintes pares *(N; K)*: *(14; 9,0)*, *(13; 9,5)* e *(12; 10,0)*, ou seja, para se obter baixas taxas de falsas aceitações e falsas rejeições, deve-se

comparar um pico no gráfico da segunda derivada da medida global de movimento com a média dos 12, 13 ou 14 pontos adjacentes (à esquerda ou à direita) daquele pico; se a amplitude do pico for 9, 9,5 ou 10 vezes maior que uma das médias calculadas dos pontos adjacentes (esquerda ou direita), então marca-se o quadro como pertencente a uma transição abrupta.

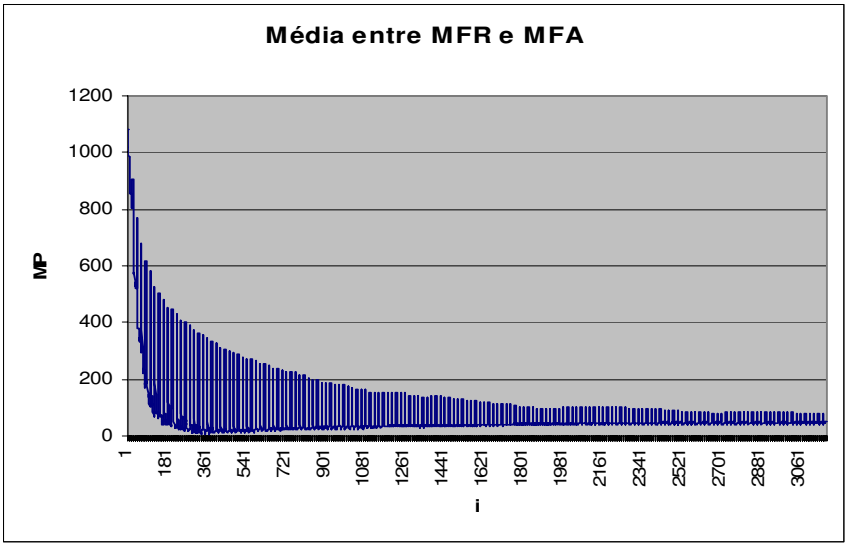


Figura 30 – Gráfico i x MP

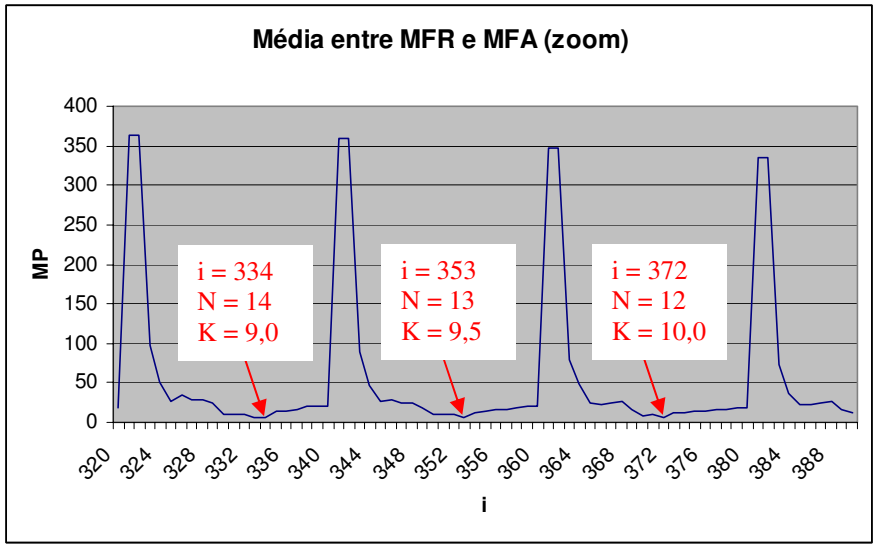


Figura 31 – Zoom do gráfico da Figura 30

Uma vez obtidos os valores otimizados de *N* e *K*, a partir dos cinco vídeos do conjunto de treinamento, foi realizado um experimento de detecção de transições abruptas utilizando-se cinco novos vídeos (conjunto de teste) e os valores 14 e 9,0, para

N e *K*, respectivamente (uma das três duplas de valores ótimos encontrados). Os quadros-chave dos vídeos do conjunto de teste podem ser encontrados no Apêndice A, Figuras 51 a 55. Vale citar que os vídeos utilizados nos experimentos, tanto no conjunto de treinamento quanto no de teste, possuem quadros com a dimensão 320 x 240 *pixels*, foram capturados nos formatos *mpeg* ou *avi*, a uma taxa de aquisição de 12 quadros por segundo, e têm um tempo de duração de aproximadamente 31 segundos cada um, o que resulta em torno de 375 quadros por vídeo. A seguir, são expostas estatísticas sobre os experimentos realizados com os dois conjuntos de vídeos, o de treinamento (Tabela 1) e o de teste (Tabela 2).

Tabela 1 – Estatísticas para o conjunto de treinamento

CONJUNTO DE TREINAMENTO							
	Vídeo 1	**Vídeo 2**	**Vídeo 3**	**Vídeo 4**	**Vídeo 5**	**Média**	**Desvio padrão**
Razão das Falsas Rejeições (%)	0	55,56	0	0	11,11	13,33	21,55
Razão das Falsas Afirmações (%)	0	0	0	0	0	0	0
Nº de transições detectadas manualmente	3	18	5	2	18	9,2	7,25
Nº de transições detectadas automaticamente	3	8	5	2	16	6,8	5,04
Nº de acertos	3	8	5	2	16	6,8	5,04
Nº de falsas rejeições	0	10	0	0	2	2,4	3,88
Nº de falsas afirmações	0	0	0	0	0	0	0

Tabela 2 – Estatísticas para o conjunto de teste

CONJUNTO DE TESTE							
	Vídeo 1	**Vídeo 2**	**Vídeo 3**	**Vídeo 4**	**Vídeo 5**	**Média**	**Desvio padrão**
Razão das Falsas Rejeições (%)	0	8,33	33,33	28,57	40	22,05	15,27
Razão das Falsas Afirmações (%)	50	16,67	0	14,29	20	20,19	16,39
Nº de transições detectadas manualmente	4	12	3	7	5	6,2	3,19
Nº de transições detectadas automaticamente	6	13	2	6	4	6,2	3,71
Nº de acertos	4	11	2	5	3	5	3,16
Nº de falsas rejeições	0	1	1	2	2	1,2	0,75
Nº de falsas afirmações	2	2	0	1	1	1,2	0,75

Como se pode ver pela Tabela 1, as taxas médias para o conjunto de treinamento foram satisfatórias (uma taxa relativamente baixa de falsas rejeições e taxa nula de

falsas afirmações). Já no que se refere ao desvio padrão, este apresentou um alto valor para as falsas rejeições, devido ao fato do Vídeo 5 e, principalmente, o Vídeo 2 terem apresentados taxas de falsas rejeições destoantes em relação aos demais vídeos, que apresentaram taxa nula de falsas rejeições. O desvio padrão para as falsas afirmações, por sua vez, foi obviamente nulo.

Analisando-se agora a Tabela 2, pode-se perceber que o experimento com o conjunto de teste apresentou taxas mais altas do que as apresentadas pelo conjunto de treinamento, tanto para as falsas rejeições quanto para as falsas afirmações. Mesmo assim, as taxas apresentadas pelo conjunto de teste são satisfatórias e promissoras. Mais adiante, serão apontadas algumas causas para a ocorrência dessas falsas rejeições e falsas afirmações, ou seja, algumas situações nas quais a estratégia para a detecção de transições abruptas proposta neste trabalho pode falhar. Um ponto que pode contribuir para a redução das taxas de falsas rejeições e de falsas afirmações segundo essa estratégia, e que é deixado como uma sugestão de trabalho futuro, é a utilização de mais vídeos no conjunto de treinamento, de modo a encontrar valores de N e K mais apropriados para uma gama maior de vídeos.

Quanto ao desvio padrão para as falsas rejeições do conjunto de teste, este foi inferior ao do conjunto de treinamento, indicando uma maior uniformidade entre os vídeos do conjunto de teste, pelo menos no que diz respeito às falsas rejeições. Já no que diz respeito ao desvio padrão para as falsas afirmações, os vídeos do conjunto de teste, ao contrário do que ocorre no conjunto de treinamento, não apresentaram apenas taxas nulas para as falsas afirmações, o que acarretou em um desvio padrão para as falsas afirmações superior (e não-nulo) no conjunto de teste, equiparável ao desvio padrão para as falsas rejeições do conjunto de teste.

Nas Figuras 32 a 37, são expostos alguns exemplos de acertos, de falsas rejeições e de falsas afirmações que ocorreram nos experimentos com os conjuntos de treinamento e de teste, utilizando-se 14 e 9,0, respectivamente, como os valores de N e K.

Figura 32 – Exemplo 1 dos experimentos em detecção de transições abruptas

transição (detecção manual) transição (detecção automática)

Figura 33 – Exemplo 2 dos experimentos em detecção de transições abruptas

transição (detecção manual)

Figura 34 – Exemplo 3 dos experimentos em detecção de transições abruptas

transição (detecção manual)

Figura 35 – Exemplo 4 dos experimentos em detecção de transições abruptas

transição (detecção automática)

Figura 36 – Exemplo 5 dos experimentos em detecção de transições abruptas

transição (detecção automática)

Figura 37 – Exemplo 6 dos experimentos em detecção de transições abruptas

Cada exemplo consiste em um pequeno trecho de um vídeo do conjunto de treinamento ou do conjunto de teste. Nos dois primeiros exemplos (Figuras 32 e 33), têm-se casos de acerto, ou seja, os quadros considerados como transições abruptas através de detecção manual (humana) também foram apontados pelo algoritmo desenvolvido neste trabalho como uma transição abrupta. Já os dois exemplos seguintes (Figuras 34 e 35) exibem casos de falsa rejeição: os quadros detectados manualmente

não constam na lista de quadros apontados pelo algoritmo como transições abruptas. Por fim, as Figuras 36 e 37 ilustram casos de falsas afirmações (o algoritmo aponta, como transições abruptas, quadros que não foram detectados manualmente).

Dentre as causas da ocorrência de falsas rejeições, ao se seguir a estratégia para a detecção de transições abruptas proposta neste trabalho, constam, por exemplo: falha humana na detecção manual (a pessoa responsável pela detecção humana classifica erroneamente um determinado quadro como uma transição abrupta; entretanto, quando o sistema acertadamente não lista tal quadro como transição abrupta, registra-se uma falsa rejeição), transição entre quadros com uma grande predominância de uma determinada intensidade (por exemplo, uma cena em que há uma símbolo vermelho em um fundo preto e em que o símbolo desaparece, restando apenas o fundo preto: o desaparecimento do símbolo vermelho é classificado pelo ser humano como uma transição abrupta; entretanto, como o algoritmo gerador de mapas de movimento mede a intensidade de movimento a partir das variações de intensidade de brilho nas regiões da cena e a cor vermelha é uma cor com baixo brilho, a mudança de um quadro preto com um símbolo vermelho para um quadro completamente preto poderia gerar baixos valores de intensidade de movimento, o que levaria o algoritmo de detecção de transições a não classificar o quadro em questão como uma transição abrupta, fato que seria computado como uma falsa rejeição) etc. A título de ilustração, é mostrado, na Figura 38, um exemplo do caso de falsa rejeição devido a uma transição entre quadros com predominância de uma determinada intensidade.

Figura 38 – Um caso típico de falsa rejeição

Na Figura 38, o fato do logotipo “apagar” (desaparecer) pode ser considerado, por um ser humano, como uma transição abrupta. Entretanto, como, nos quadros em que aparece o logotipo, há uma predominância de uma cor escura (a cor vermelha) e os quadros a seguir são completamente escuros, o algoritmo não detecta uma grande disparidade entre os quadros. Desse modo, o algoritmo considera que, nesses quadros, há um movimento de baixa intensidade e não detecta nenhuma transição abrupta, o que é contabilizado como uma falsa rejeição.

Já no que se refere a possíveis causas da ocorrência de falsas afirmações, pode-se citar falha humana na detecção manual (de maneira oposta ao que ocorre no caso de falsas rejeições), o surgimento de legendas, anotações, logotipos e outros símbolos textuais ou gráficos nas cenas (que, muitas vezes, não é considerado pelo observador como uma transição abrupta, mas que pode gerar altos valores de disparidade entre os quadros, resultando em altas intensidades de movimento e classificação de tais quadros como transições abruptas pela estratégia proposta neste trabalho) etc. Vale citar que a Figura 37 ilustra um caso de falsa afirmação por parte do algoritmo devido ao surgimento de uma legenda na cena.

O algoritmo desenvolvido neste trabalho e comentado nesta seção constitui-se como uma forma simples de se detectar transições abruptas em seqüências de vídeos. Tal algoritmo se utiliza das informações obtidas pelo algoritmo gerador de mapas de movimento (também desenvolvido neste trabalho e detalhado na Seção 4.2) e realiza, quando da utilização de uma das três duplas ótimas de parâmetros N e K (determinadas experimentalmente), uma detecção de transições abruptas com percentuais médios relativamente baixos de falsas rejeições e de falsas afirmações.

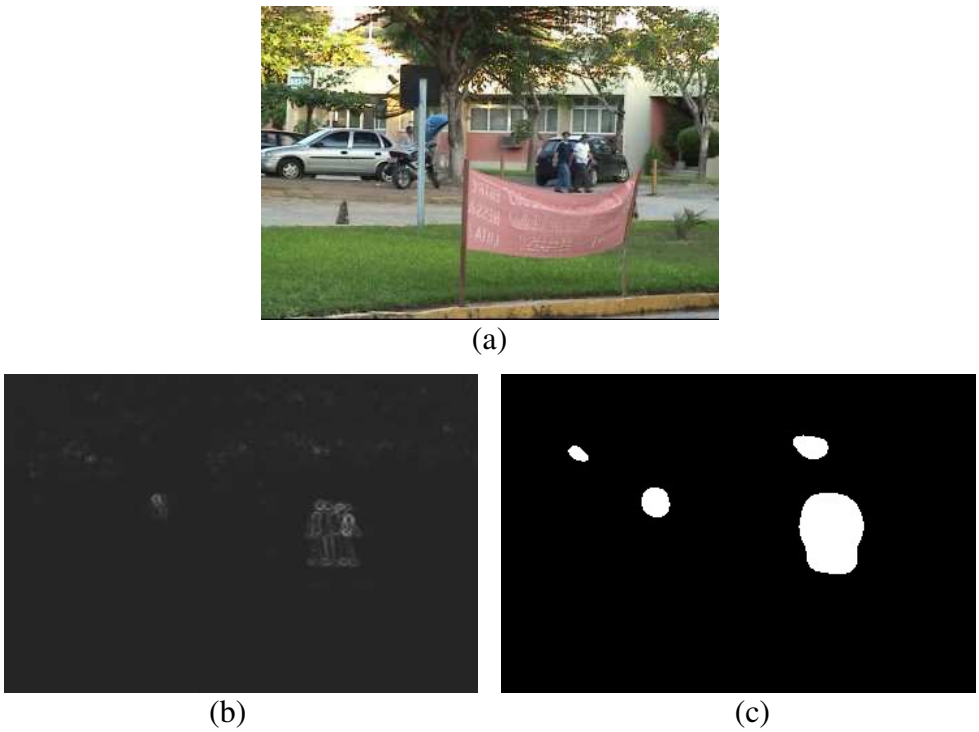
5.3 Segmentação de Objetos Móveis

Na Seção 4.1, foi apresentado o Módulo de Segmentação de Movimento, um dos módulos que compõem o sistema de Atenção Visual implementado neste trabalho. Também foi comentado que tal módulo já se constitui, por si só, como uma contribuição deste trabalho. O Módulo de Segmentação de Movimento recebe um vídeo A e um vídeo M (composto pelos mapas de movimento referentes aos quadros do vídeo A, sendo estes mapas gerados pelo algoritmo gerador de mapas de movimento desenvolvido neste trabalho e exposto na Seção 4.2) e gera como resposta um vídeo S, cujos quadros são idênticos aos quadros do vídeo A, só que com as regiões estáticas completamente em preto (as informações sobre quais são as regiões estáticas e dinâmicas são obtidas a partir do vídeo M). O efeito desse processo é a segmentação dos objetos móveis do vídeo A (e de uma pequena área em torno deles).

O procedimento para gerar o vídeo S é bastante simples. Como pode ser visto na Seção 5.1, os mapas de movimento apresentam o contorno dos objetos móveis. Ao se aplicar uma máscara Gaussiana de valor alto e se efetuar uma binarização com um

limiar baixo em cada um dos mapas de movimento, os objetos móveis passam a se apresentar como áreas fechadas (que abrangem os objetos móveis e uma pequena região em torno deles) de intensidade máxima (255), enquanto as demais porções dos mapas se apresentam completamente escuras (intensidade 0). A aplicação de uma máscara Gaussiana alta gera o efeito de "borrão" nos contornos dos objetos móveis existentes nos mapas de movimento, enquanto o uso de um limiar baixo na binarização (*pixels* abaixo do limiar recebem intensidade 0, e *pixels* com valores iguais ou superiores ao limiar recebem intensidade 255) faz com que a maior parte dos *pixels* gerados pela aplicação da máscara Gaussiana sejam capturados. Foram utilizados, neste trabalho, uma máscara Gaussiana 11x11 e um limiar de binarização igual a 2% da máxima intensidade do mapa de movimento (valores determinados experimentalmente).

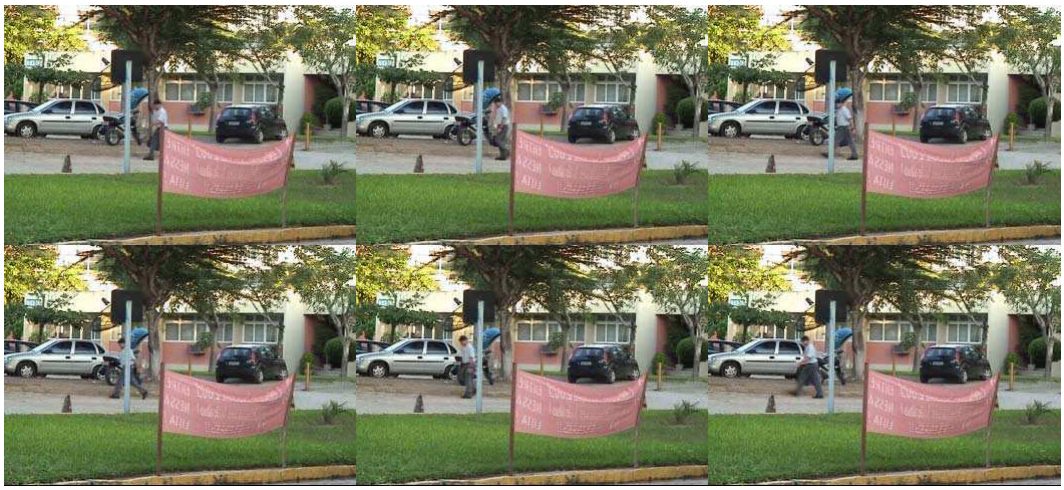
Na Figura 39, é apresentado um exemplo da transformação de um mapa de movimento (que apresenta os contornos dos objetos móveis) em uma imagem com áreas fechadas, indicando as regiões onde ocorrem o movimento. As Figuras 39a, 39b e 39c apresentam, respectivamente, um quadro original de um vídeo, seu mapa de movimento (gerado pelo Módulo de Atenção Temporal) e uma imagem com as regiões de movimento (mapa de movimento processado), gerada a partir do mapa de movimento.



(a) (b) (c)

Figura 39 – Exemplo de geração de um mapa de movimento processado: (a) quadro original; (b) mapa de movimento e (c) mapa de movimento processado

Com esses mapas de movimento processados em mãos, basta efetuar as seguintes alterações em cada um dos quadros do vídeo *A*: as regiões do quadro que correspondem a regiões pretas no mapa de movimento processado específico são "pintadas" de preto. As demais regiões permanecem inalteradas. As Figuras 40 e 41 exibem alguns exemplos do processamento realizado pelo Módulo de Segmentação de Movimento. Em cada exemplo, são mostrados, primeiramente, alguns quadros de um vídeo (um "vídeo *A*") e, em seguida, os quadros segmentados pelo movimento correspondentes (quadros do "vídeo *S*").

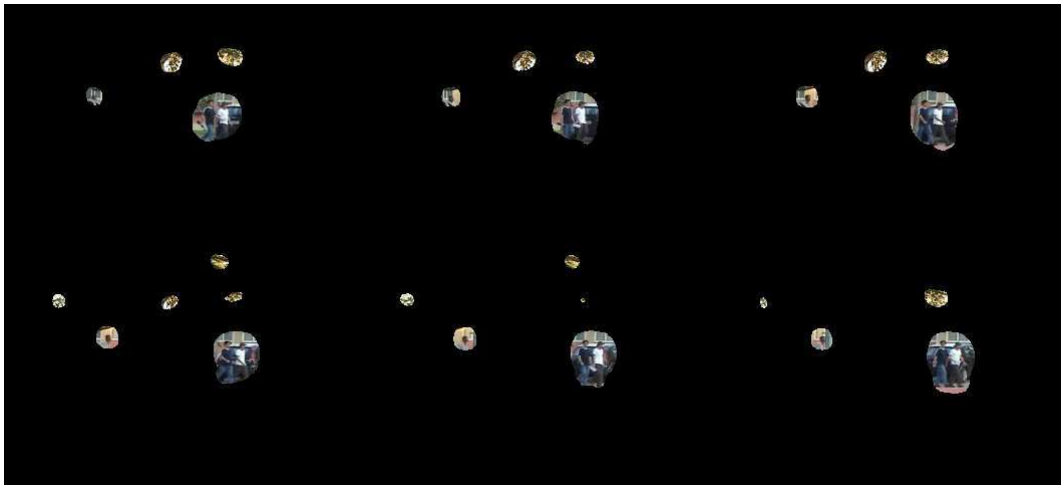


(a)

(b)

Figura 40 – Exemplo 1 de segmentação de movimento: (a) quadros de entrada; (b) quadros segmentados (saída)

(a)



(b)

Figura 41 – Exemplo 2 de segmentação de movimento: (a) quadros de entrada; (b) quadros segmentados (saída)

Ao se comparar os quadros segmentados com seus respectivos quadros originais nas Figuras 40 e 41, pode-se perceber que, nos quadros segmentados, realmente estão visualizáveis apenas os objetos que estão em movimento na cena visual (no caso de ambos exemplos dados, pessoas caminhando e as extremidades de galhos de árvore balançando devido ao vento) e uma pequena área ao seu redor. Todo o restante da cena está completamente escuro. Observando-se os quadros segmentados, pode-se acompanhar (identificar quadro a quadro) facilmente os objetos móveis (seja manual ou automaticamente), que são o objeto de interesse no Módulo de Segmentação de Movimento. Quanto aos distratores, sua proporção é reduzida drasticamente, devido à eliminação das regiões estáticas. No caso dos exemplos dados, os únicos distratores remanescentes são as extremidades dos galhos das árvores, que não constituem, nas

cenas analisadas, objetos interessantes, mas que não são excluídas dos quadros segmentados por apresentarem movimento significativo.

Em um contexto em que se deseja realizar uma análise mais profunda apenas nas regiões do vídeo onde ocorre movimento (por serem consideradas maiores merecedoras de atenção), o sistema desempenha tal funcionalidade de forma simples e efetiva: ao eliminar as regiões estáticas da cena (deixá-las completamente escuras), o sistema reduz enormemente a quantidade de informações que precisarão ser analisadas em cada quadro do vídeo. Nesse contexto, foi realizado um experimento trivial de contagem de *pixels*. Em cenas como as mostradas nas Figuras 40 e 41, em que há apenas o movimento dos galhos das árvores e de poucas pessoas caminhando a uma certa distância da câmera, a redução da quantidade de *pixels* a serem analisados (o percentual de *pixels* completamente escuros), em média, ultrapassa os 97%, o que se configura como uma taxa bastante promissora. Durante o vídeo, o percentual de *pixels* completamente escuros varia bastante, a depender do tipo de cena que está sendo analisada. A taxa de 97% se restringe a cenas como as mostradas nas Figuras 40 e 41. Em outras cenas, essa taxa sofre uma redução (por exemplo, em cenas em que o objeto móvel passa bem próximo à câmera) ou um aumento (por exemplo, em cenas em que os únicos objetos móveis são as extremidades dos galhos das árvores balançando).

Como um trabalho futuro, pretende-se realizar uma avaliação numérica de desempenho na detecção de objetos em movimento. Tal avaliação consistiria em rotular manualmente as regiões de cada quadro dos vídeos quanto à ocorrência ou não de movimento. Em seguida, essa detecção manual seria comparada à detecção realizada pelo sistema e seriam calculadas estatísticas como as realizadas no caso da detecção de transições abruptas (taxas de falsas rejeições e de falsas afirmações, entre outras).

5.4 Integração entre Atenção Espacial *bottom-up* e Atenção Temporal

Esta seção discute os experimentos para a geração, a partir dos quadros de um vídeo, de um novo vídeo, composto pelos mapas de saliência segmentados dos quadros do vídeo original. Como citado na Seção 4.1, um mapa de saliência segmentado foi o nome dado neste trabalho a um quadro de um vídeo modificado da seguinte forma: (i) as regiões estáticas são apresentadas completamente em preto (com intensidades nulas)

e (ii) ao invés das regiões da imagem onde ocorre movimento são exibidas as características espaciais *bottom-up* dessas regiões. A seguir, um resumo do procedimento para a formação do mapa de saliência segmentado (Seção 4.1): o Módulo de Segmentação de Movimento recebe o quadro do vídeo e seu mapa de movimento (fornecido pelo Módulo de Atenção Temporal) e gera um quadro segmentado pelo movimento. O Módulo de Atenção Espacial *bottom-up*, por sua vez, recebe esse quadro segmentado e gera um mapa de saliência *bottom-up*. Obviamente, as regiões pretas presentes no quadro segmentado pelo movimento (correspondentes a regiões estáticas no quadro original) permanecerão idênticas no mapa de saliência *bottom-up*. É esse mapa de saliência *bottom-up* que recebe, neste trabalho, a alcunha de mapa de saliência segmentado.

Para a geração do mapa de saliência *bottom-up* (mapa de saliência segmentado) a partir do quadro segmentado, foram realizadas algumas pequenas adaptações no programa desenvolvido e cedido por Rodrigues [Rodrigues, 2002]. Tal programa se baseia no modelo de Atenção Espacial *bottom-up* proposto por Itti [Itti, 2003; Itti and Koch, 2001; Itti, 2000; Itti et al., 1998] e utiliza intensidade, cor e orientação como as características visuais primitivas na geração do mapa de saliência *bottom-up*. Seguem alguns exemplos de experimentos realizados com o programa de Rodrigues [Rodrigues, 2002].

Figura 42 – Exemplo 1 de Atenção Espacial *bottom-up*: imagem original e mapa de saliência *bottom-up* correspondente

Figura 43 – Exemplo 2 de Atenção Espacial *bottom-up*: imagem original e mapa de saliência *bottom-up* correspondente

Rodrigues [Rodrigues, 2002] trabalhou apenas com as evidências espaciais *bottom-up* das cenas visuais (imagens estáticas). Tais evidências foram utilizadas para a localização automática de placas de sinalização. Em seu trabalho, foram aplicados pesos iguais para cada uma das três características (intensidade, cor e orientação) utilizadas na composição do mapa *bottom-up*.

Do mesmo modo, no trabalho apresentado nesta dissertação, tanto nos exemplos das Figuras 42 e 43, quanto na geração dos mapas de saliência segmentados, também foram utilizados intensidade, cor e orientação como as características visuais primitivas, e pesos iguais para a composição dessas características. Entretanto, ao contrário do trabalho de Rodrigues [Rodrigues, 2002], o objetivo aqui é trabalhar com seqüências de vídeo e não com imagens estáticas. Assim sendo, o resultado final do sistema proposto neste trabalho consta não no mapa de saliência *bottom-up* de uma determinada imagem, e sim em um vídeo completo, em que os quadros são mapas de saliência *bottom-up* com as regiões estáticas completamente em preto (integração de evidências espaciais e temporais). Desse modo, no vídeo final, são exibidas as informações *bottom-up* apenas dos objetos móveis.

Nas Figuras 44 e 45, são exibidos alguns exemplos dos resultados finais produzidos pelo sistema proposto neste trabalho. Cada figura apresenta um pequeno trecho de vídeo gerado por esse sistema.

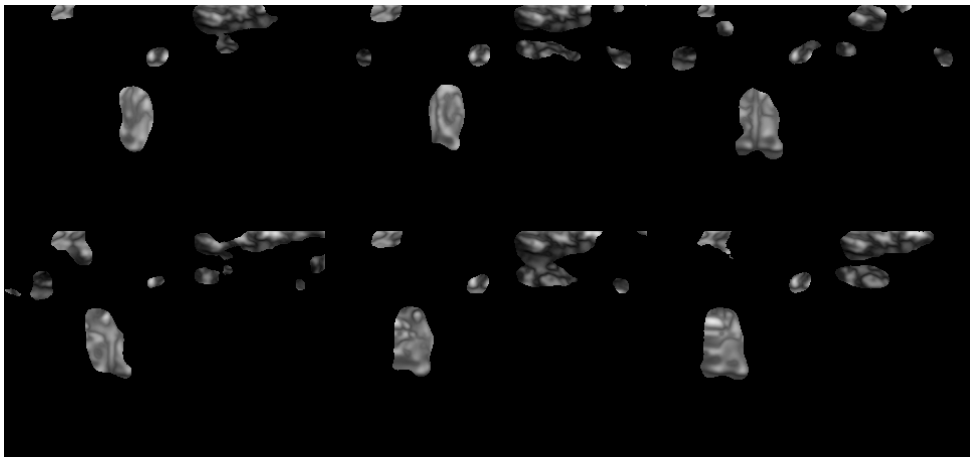


Figura 44 – Exemplo 1 de integração entre Atenção Temporal e Atenção Espacial *bottom-up*: quadros de saída

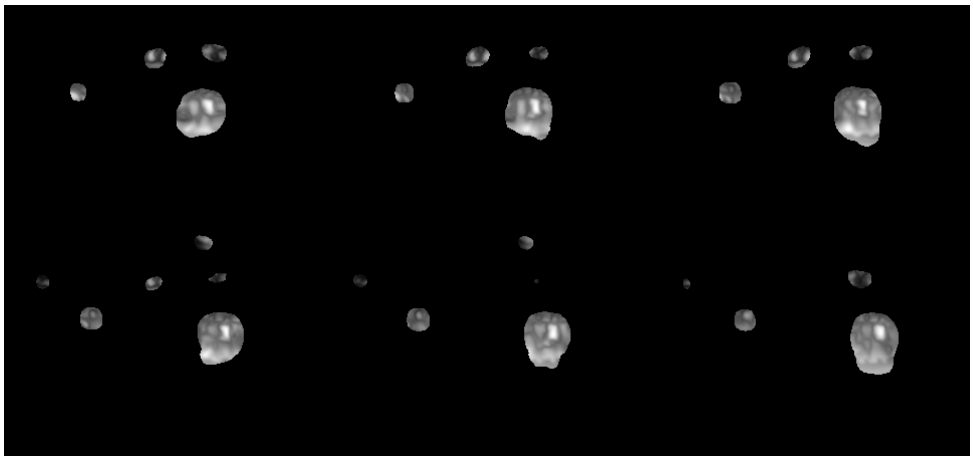


Figura 45 – Exemplo 2 de integração entre Atenção Temporal e Atenção Espacial *bottom-up*: quadros de saída

Os trechos dos vídeos originais correspondentes aos trechos expostos nas Figuras 44 e 45 foram expostos na Seção 5.3 (respectivamente, Figuras 40a e 41a). Naquela ocasião, esses mesmos quadros de vídeo foram selecionados como ilustração da entrada do Módulo de Segmentação de Movimento. As Figuras 40a e 41a reproduzem novamente amostras desses quadros de entrada e as Figuras 44 e 45 apresentam os resultados da integração da Atenção Temporal com a Atenção Espacial *bottom-up*, para os quadros das Figuras 40a e 41a, respectivamente.

Assim como no trabalho de Rodrigues [Rodrigues, 2002], em que o sistema de Atenção Espacial *bottom-up* foi utilizado para se localizar placas de sinalização (objetos que apresentam, normalmente, saliência *bottom-up* superior aos demais objetos presentes na mesma cena visual), o sistema proposto neste trabalho (Atenção Temporal + Atenção Espacial *bottom-up*) pode ser aplicado para se detectar regiões com alta saliência *bottom-up* em objetos móveis. Uma possível abordagem dentro dessa linha seria, por exemplo, localizar placas de veículos em movimento, entre outras.

Como já citado, os experimentos realizados neste trabalho utilizaram um mesmo peso para a intensidade, a cor e a orientação. Entretanto, esses pesos poderiam ser facilmente alterados, para se obterem melhores resultados, a depender do tipo de problema que se esteja tentando solucionar. No caso citado (localizar placas de veículos em movimento), por exemplo, talvez fosse mais interessante colocar um peso maior na característica orientação, em detrimento das demais.

Outro exemplo de aplicação para os resultados finais de um sistema como o desenvolvido neste trabalho é a detecção de faces de pessoas que estão em movimento.

O fato de o mapa de saliência segmentado não apresentar as regiões estáticas das cenas filtra a quantidade de regiões onde se deve procurar uma face humana e descarta as faces das pessoas que não estão se movimentando. Além disso, como certas regiões da face humana (os olhos, por exemplo) possuem grande saliência *bottom-up*, as características *bottom-up* apresentadas pelo mapa de saliência segmentado facilitam a detecção de regiões da face, pois, ao se encontrar uma determinada região da face (um olho, por exemplo), e de posse do conhecimento da anatomia facial (posição relativa entre os olhos, o nariz, a boca e demais regiões da face), a tarefa de determinar outras regiões da face e a face como um todo se torna bem mais simples.

Dentre outras possíveis aplicações, pode-se citar um sistema de monitoramento de vídeo (*video surveillance*), problemas de indexação de vídeo (em empresas de televisão), bancos de dados multimídia etc.

5.5 Detalhes de Implementação

Esta seção visa expor detalhes de implementação a respeito dos experimentos realizados neste trabalho. Primeiramente, no que diz respeito ao ambiente operacional, todos os experimentos foram implementados e testados no sistema operacional *Linux*. Já no que se refere às características médias dos computadores utilizados, pode-se citar processador com *clock* de 2 GHz, memória de 512 MB e *hard disk* IDE UDMA 133.

Passando para os detalhes sobre as ferramentas utilizadas, a extração dos quadros dos vídeos originais (utilizados como entrada de diversos experimentos realizados neste trabalho) e a geração de seqüências de vídeo a partir da composição dos quadros gerados como saída pelos módulos do sistema de Atenção Visual desenvolvido neste trabalho (mapas de movimento, mapas de movimento processados, quadros segmentados pelo movimento e mapas de saliência *bottom-up* segmentados pelo movimento) foram realizadas no sistema operacional *Windows*, com a utilização do *software* Adobe Premiere [Adobe Premiere].

Já no quesito programação, foram utilizados a linguagem C/C++ e os pacotes de bibliotecas *IPP* (*Integrated Performance Primitives* – primitivas de desempenho integrado) e *OpenCV* (*Open source Computer Vision library* – biblioteca de Visão Computacional com código aberto), ambos fornecidos pela Intel e voltados para a linguagem de programação C/C++. O *IPP* [IPP; INTEL] é um pacote para

processamento de sinais, processamento de imagens e cálculo de matrizes, e que oferece uma gama de funções de baixo-nível. Já o *OpenCV* [OpenCVa; OpenCVb] é um pacote que fornece funções de alto nível para Visão Computacional e Processamento de Imagens.

Finalmente, a Tabela 3 apresenta um sumário do desempenho médio dos Módulos de Atenção Temporal, de Segmentação de Movimento e de Atenção Espacial *bottom-up*. Vale frisar que, para se calcular a velocidade média de processamento para cada um desses módulos, foram utilizadas imagens com a dimensão 352 x 240 *pixels* (dimensão dos quadros dos vídeos originais nos experimentos com o sistema de Atenção Visual desenvolvido neste trabalho).

Tabela 3 – Desempenho médio dos módulos do sistema de Atenção Visual

	Velocidade de processamento (quadros/segundo)
Módulo de Atenção Temporal	12,422
Módulo de Segmentação de Movimento	1,613
Módulo de Atenção Espacial *bottom-up*	0,012

Como se pode perceber pela Tabela 3, o desempenho médio varia bastante entre os diferentes módulos. Embora questões como velocidade de processamento não sejam o objetivo final deste trabalho, o algoritmo para a geração de mapas de movimento (utilizado no Módulo de Atenção Temporal) apresentou um desempenho médio satisfatório, apesar de insuficiente para operar em tempo real. O Módulo de Segmentação de Movimento, por sua vez, apresentou um desempenho bem inferior ao do Módulo de Atenção Temporal, mas ainda aceitável para processamentos *off-line*. Por fim, o programa desenvolvido por Rodrigues [Rodrigues, 2002] para a extração de características espaciais *bottom-up*, e cedido pelo autor para incorporação no sistema desenvolvido neste trabalho (sob a forma do Módulo de Atenção Espacial *bottom-up*) também não objetivava desempenho computacional e apresentou uma velocidade média de processamento bastante baixa. Possíveis trabalhos futuros consistiriam em aperfeiçoar os processamentos realizados por esses três módulos, principalmente pelo Módulo de Atenção *bottom-up*, de modo a aumentar seus desempenhos computacionais médios.

Capítulo 6

Conclusões e Sugestões para Trabalhos Futuros

Nos últimos anos, a Atenção Visual tem se destacado como uma área de pesquisa promissora, seja em sua vertente espacial (já bastante explorada) ou temporal (vertente mais recente). Neste trabalho, o foco principal recaiu exatamente na proposição e implementação de um modelo de Atenção Visual que integre essas duas vertentes (particularmente, as vertentes espacial *bottom-up* e temporal).

Baseando-se no algoritmo proposto por Wildes [Wildes, 1998], foi desenvolvido um algoritmo que recebe os quadros de um vídeo e gera como resultado seus respectivos mapas de movimento (imagens em tom de cinza que representam a intensidade de movimento na cena). A partir das evidências temporais fornecidas por esses mapas de movimento, foi desenvolvida uma estratégia para a detecção de transições abruptas em vídeo. Dado um vídeo, os quadros que representavam transições abruptas eram selecionados primeiro manualmente (detecção humana) e depois automaticamente (através da aplicação da estratégia proposta neste trabalho). A comparação dos resultados das detecções manual e automática serviu de métrica para avaliar a estratégia proposta, estratégia essa que apresentou um desempenho promissor, gerando, ao se utilizar determinados parâmetros (encontrados experimentalmente e considerados, neste trabalho, como parâmetros ótimos para o algoritmo proposto), percentuais relativamente baixos de falsas rejeições e de falsas afirmações.

A seguir, entraram em cena os experimentos em segmentação de movimento, sob a forma de um algoritmo que recebe os mapas de movimento gerados pelo algoritmo desenvolvido neste trabalho (são extraídas, dos mapas de movimento, informações sobre quais são as regiões estáticas e as regiões dinâmicas da imagem) e os quadros do vídeo original e gera um novo vídeo, cujos quadros são os do vídeo original

segmentados pelo movimento (os quadros originais com as regiões estáticas completamente em preto). Nesse vídeo se tem a impressão de focos de luz que acompanham os objetos móveis, iluminando a eles e a uma pequena área ao seu redor, enquanto todo o restante da cena visual está na mais completa e total "escuridão".

No que se refere ao sistema que integra Atenção Espacial *bottom-up* e Atenção Temporal, ele foi dividido em alguns módulos, de forma a se atingir, passo a passo, os resultados desejados. O algoritmo gerador de mapas de movimento e o algoritmo de segmentação de movimento, ambos já citados anteriormente, constituem dois de seus módulos: o de Atenção Temporal e o de Segmentação de Movimento. Desse modo, para se efetuar a integração das Atenções Espacial e Temporal segundo o modelo proposto neste trabalho, deve-se continuar o procedimento a partir do ponto em que o Módulo de Segmentação de Movimento parou.

O vídeo com os quadros segmentados pelo movimento são, então, enviados ao Módulo de Atenção Espacial *bottom-up*. Esse módulo recebe os quadros de um vídeo e retorna os mapas de saliência *bottom-up* referentes a esses quadros. Como os quadros recebidos pelo Módulo de Atenção Espacial *bottom-up* estão segmentados pelo movimento, os mapas de saliência *bottom-up* gerados também possuem as regiões estáticas completamente em preto e receberam, neste trabalho, o nome de mapas de saliência segmentados. A obtenção do vídeo composto por esses mapas segmentados determina o fim da execução do sistema de integração das Atenções Espacial e Temporal. Tal integração se concretiza no fato de que, no vídeo final, as regiões estáticas são completamente descartadas (são representadas em preto) e são exibidas as características espaciais *bottom-up* apenas das regiões onde ocorre movimento.

No que se refere aos resultados obtidos, pode-se citar o sistema (módulo) de Atenção Temporal, que se configura como um algoritmo simples que realiza uma detecção versátil do movimento presente em cenas de vídeo. Já os experimentos com detecção de transições abruptas manifestaram resultados bastante promissores, como, por exemplo, taxas relativamente baixas de falsas rejeições e de falsas afirmações. Quanto ao Módulo de Segmentação de Movimento, os resultados apresentados, apesar de terem sido avaliados subjetivamente, apontam na direção de uma ótima taxa na detecção dos objetos móveis em vídeos. Finalmente, em relação à integração entre Atenção Temporal e Espacial, uma vez que a extração das características espaciais *bottom-up* só ocorre nas regiões em que o movimento foi detectado (saída do Módulo

de Segmentação de Movimento), a abordagem proposta permite a determinação das regiões mais salientes (do ponto de vista estático e dinâmico) de uma forma mais eficaz.

Um trabalho futuro interessante, de forma a dar continuidade ao trabalho aqui desenvolvido, seria testar a geração dos mapas de saliência segmentados com diferentes pesos para as características visuais primitivas da Atenção Espacial *bottom-up*, de modo a se determinar quais seriam as melhores combinações para determinados tipos de problema. Outra sugestão consistiria na implementação e integração do Módulo de Visão Estéreo (que está representado no modelo de Atenção Visual proposto neste trabalho, mas que não foi implementado), para uma especificidade ainda maior do tipo de conteúdo que se deseja analisar (conteúdo sobre o qual deve recair o foco de atenção, dependente do tipo de aplicação).

Persistindo na idéia de se obter mais evidências sobre as cenas analisadas (além das evidências temporais e espaciais *bottom-up*, a partir do sistema desenvolvido neste trabalho, e de evidências de profundidade, sugeridas como um trabalho futuro), para se realizar uma melhor detecção das regiões merecedoras de atenção, um possível trabalho futuro seria a integração de evidências espaciais *top-down* ao modelo aqui proposto e a conseqüente implementação de um Módulo de Atenção Espacial *top-down*.

Já no que se refere aos experimentos em detecção de transições abruptas, considerando que a curva do comportamento do movimento pode ser vista como um sinal temporal geral, pode-se imaginar a utilização de métodos bem estabelecidos de análise de sinais no domínio do tempo, a exemplo dos métodos para processamento de voz, como a *AMDF* (*Average Magnitude Difference Function* – função da diferença média das magnitudes) [Ross et al., 1974] e métodos para a determinação da freqüência fundamental [Gold and Morgan, 2000; Rabiner and Juang, 1993; Rabiner and Schafer, 1978]. Os resultados de um processamento com esses métodos poderiam então ser utilizados para auxiliar na caracterização do tipo de vídeo que está sendo analisado (por exemplo, contendo muita ou pouca ação, suave etc).

Quanto à detecção de movimento, esta foi realizada, neste trabalho, a partir da análise unicamente de pares de quadros consecutivos dos vídeos. Uma outra possibilidade, de modo a aprimorar tal processo e se conseguir uma certa invariância com relação à taxa de aquisição do vídeo, seria a análise de quadros não necessariamente adjacentes (seriam analisados pares de quadros distanciados de k, com $k = 1, 2, 3, ...$).

Por fim, pode-se citar mais algumas sugestões para trabalhos futuros, como, por exemplo, um detector de outros tipos de transições (não necessariamente abruptas), um sistema de *video surveillance* semi-automático (que, em tempo real, mostrasse apenas as regiões do vídeo onde ocorre movimento) etc.

Referências Bibliográficas

[Adobe Premiere] Adobe Premiere Software.
http://www.adobe.com/products/premiere/main.html. Último acesso: 12/08/2005.

[Andersen et al., 1990] Andersen, R. A.; Bracewell, R. M.; Barash, S.; Gnadt, J. W. and Fogassi, L. Eye position effects on visual, memory, and saccade-related activity in areas lip and 7a of macaque. *Journal of Neuroscience*, vol. 10, pp. 1176-1196, 1990.

[Ballard and Brown, 1982] Ballard, D. H. and Brown, C. M. *Computer Vision*. Prentice-Hall, Englewood Cliffs, New Jersey, 1982.

[Barron et al., 1994] Barron, J. L. et al. Performance of Optical Flow Techniques. *In: International Journal of Computer Vision*, vol.12, no.1, pp.43-77, 1994.

[Bauchspiess, 2002] Bauchspiess, Adolfo. Sistemas Inteligentes: redes neurais e lógica fuzzy. *In: VI Semana de Engenharia Elétrica da Universidade de Brasília*, Brasília, dezembro 2002.

[Beauchemin and Barron, 1995] Beauchemin, S. S. and Barron, J. L. The Computation of Optical Flow. *Journal Surveys*, vol. 27, no. 3, September 1995.

[Burt and Adelson, 1983] Burt, P. J. and Adelson, E. H. The Laplacian Pyramid as a Compact Image Code. *In: IEEE Transactions on Communications*, COM-31(4), pp. 532–540, 1983.

[Capurro et al., 1997] Capurro, C.; Panerai, F. and Sandini G. Dynamic Vergence using Log-polar Images. *International Journal of Computer Vision*, vol. 24, no. 1, pp. 79-94, Aug. 1997.

[Carpenter et al., 1998] Carpenter, G.; Grossberg, S. and Lesher, G. The Representation of Visual Salience in Monkey Parietal Cortex. *Nature*, vol. 391, pp. 481-484, 1998.

[Chen et al., 2003] Chen, Li-Qun; Xie, Xing; Fan, Xin; Ma, Wei-Ying; Zhang, Hong-Jiang and Zhou, He-Qin. A visual attention model for adapting images on small displays. *ACM Multimedia Systems Journal*, vol. 9, no. 4, Oct. 2003.

[Chung et al., 1999] Chung, M. G. et al. Automatic video segmentation based on spatio-temporal features. *Korea Telecom Journal*, vol. 4, no. 1, pp. 4-14, 1999.

[Collins et al., 2000] Collins, R. T. et al. *A System for Video Surveillance and Monitoring*. CMU-RI-TR-00-12, Carnegie Mellon University, May 2000.

[Coull and Nobre, 1998] Coull, J. T. and Nobre, A. C. Where and When to Pay Attention: the neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI. *Journal of Neuroscience*, vol. 18, no.18, pp.7426-7435, September 15, 1998.

[Desimone and Duncan, 1995] Desimone, R. and Duncan, J. Neural Mechanisms of Selective Visual Attention. *Nature Reviews Neuroscience*, vol. 18, pp. 193-222, 1995.

[Duncan et al., 1997] Duncan, J. et al. Integrated Mechanisms of Selective Attention. *Current Opinion in Biology*, vol. 7, pp. 255-261, 1997.

[Gilbert and Wiesel, 1989] Gilbert, C. D. and Wiesel, T. N. Columnar Specificity of Intrinsic Horizontal and Corticocortical Connections in Cat Visual Cortex. *Nature Reviews Neuroscience*, vol. 9, pp. 2432-2442, 1989.

[Gold and Morgan, 2000] Gold, B. and Morgan, N. *Speech and Audio Signal Processing*. John Wiley and Sons, 2000.

[Gomes and Fisher, 2003] Gomes, H. M. and Fisher, R. B. Primal-Sketch Feature Extraction from a Log-Polar Image. *Pattern Recognition Letters*, vol. 24, no. 7, pp. 983-992, Holanda, 2003.

[Gomes and Fisher, 2001] Gomes, H. M. and Fisher, R. B. Learning and Extracting Primal-Sketch Features in a Log-Polar Image Representation. *In: Proceedings of XIV Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, 2001, Florianópolis. IEEE Computer Society Press, pp. 338-345, 2001.

[Gomes et al., 1998] Gomes, H. M.; Fisher, R. B. and Hallam, J. A. Retina-Like Image Representation of Primal-Sketch Features Extracted Using a Neural Network Approach. *In: Proceedings of Noblesse Workshop on non-Linear Model Based Image Analysis*, 1998, Glasgow. Springer-Verlag, pp. 251-256, 1998.

[Gonçalves, 1999] Gonçalves, S. E. *Reconhecimento Visual Atencional*. PhD thesis, COPPE/UFRJ, D.Sc., Engenharia de Sistemas e Computação, 1999.

[Guimarães et al., 2001] Guimarães, S. J. F.; Couprie, M.; Leite, N. J. and Araújo, A. A. A method for cut detection based on visual rhythm. *In: Proceedings of the Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, pp. 297-304, Brazil, 2001. ISBN 0769513301.

[Horn, 1986] Horn, B. K. P. *Robot Vision*. MIT Press, Cambridge, Massachusetts, 1996.

[Humphreys, 1998] Humphreys, G. W. Neural Representation of Objects in Space: a dual coding account. *Philosophical Transactions of the Royal Society*, B353, pp. 1341-1351, 1998.

[INTEL] INTEL Corporation, http://www.intel.com. Último acesso: 10/08/2005.

[IPP] IPP - Integrated performance primitives for Intel architecture - reference manual, INTEL Corporation, 2000-2001.
http://www.intel.com/design/strong/manuals/278288.htm. Último acesso: 10/08/2005.

[Itti, 2003] Itti, L. Visual Attention. *In: The Handbook of Brain Theory and Neural Networks, 2nd Ed.,* (M. A. Arbib Ed.), pp. 1196-1201, MIT Press, Jan. 2003.

[Itti, 2000] Itti, L. *Models of Bottom-Up and Top-Down Visual Attention*. PhD thesis, California Institute of Technology, 2000.

[Itti and Koch, 2001] Itti, L. and Koch, C. Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*, 2(3), pp.194–203, 2001.

[Itti et al., 1998] Itti, L.; Koch, C. and Niebur, E. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *In: IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.20, no. 11, pp.1254–1259, 1998.

[Jaraba et al., 2003] Jaraba, E. H.; Urunuela, C. O. and Senar J. Detected Motion Classification with a Double-background and a Neighbourhood-based Difference. *Pattern Recognition Letters*, Elsevier, Location, pp. 2079-82, 2003(24).

[Koch and Ullman, 1985] Koch, C. and Ullman, S. Shifts in Selective Visual Attention: towards the underlying neural circuitry. *Human Neurobiology*, vol.4, pp. 219-227, 1985.

[Logothetis et al., 1995] Logothetis, N. K.; Pauls, J. and Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, vol. 5, pp. 552-563, 1995.

[Ma and Zhang, 2001] Ma, Y. F. and Zhang, H. J. A New Perceived Motion-based Shot Content Representation. *In: Proceedings of International Conference on Image Processing*, 2001.

[Machado, 1994] Machado, Alexei Manso Corrêa. *Metodologias para Reconhecimento de Padrões em Visão Computacional*. Dissertação de Mestrado, Universidade Federal de Minas Gerais, 1994.

[Maki et al., 2000] Maki, A.; Nordlund, P. and Eklundh J. Attentional Scene Segmentation: integrating depth and motion. *Computer Vision and Image Understanding*, vol. 78, no. 3, pp. 351-373, 2000.

[Marr, 1982] Marr, D. *Vision*. W. H. Freeman & Company, New York, 1982.

[MRTStereo] MRTStereo Project.
http://www-student.informatik.uni-bonn.de/~gerdes/MRTStereo/.
Último acesso: 12/08/2005.

[Negadharipour, 1998] Negadharipour, S. Revised Definition of Optical Flow: integration of radio-metric and geometric cues for dynamic scene analysis. *In: IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, pp. 961-979, 1998.

[Niebur and Koch, 1998] Niebur, E. and Koch C. Computational Architectures for Attention. R. Parasuraman, *The Attentive Brain*, pp. 163-186. MIT Press, Cambridge, Massachusetts, 1998.

[Northdurft, 1990] Nothdurft, H. C. Texture discrimination by cells in the cat lateral geniculate nucleus. *Experimental Brain Research*, vol. 82, pp. 48-66, 1990.

[OpenCVa] Open source computer vision library - reference manual, INTEL Corporation, 1999-2001.
http://www.cs.unc.edu/Research/stc/FAQs/OpenCV/OpenCVReferenceManual.pdf.
Último acesso: 10/08/2005.

[OpenCVb] OpenCV Project, http://sourceforge.net/projects/opencvlibrary. Último acesso: 10/08/2005.

[Ouerhani, 2004] Ouerhani, Nabil. *Visual Attention: From bio-inspired modeling to real-time implementation*. PhD thesis, University of Neuchâtel, Jan 2004.

[Pasupathy and Connor, 1999] Pasupathy, A. and Connor, C. E. Responses to contour features in macaque area v4. *Journal of Neurophysiology*, vol. 82, pp. 2490-2502, 1999.

[Posner et al., 1980] Posner, M. I.; Snyder C. R. and Davidson B. J. Attention and the Detection of Signals. *Journal of Experimental Psychology: General*, vol. 109, pp. 160-174, 1980.

[Rabiner and Juang, 1993] Rabiner, L. and Juang, B. H. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[Rabiner and Schafer, 1978] Rabiner, L. R. and Schafer, R. W. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

[Rao and Ballard, 1997] Rao, R. P. N. and Ballard, D. H. *Localized receptive fields may mediate transformation-invariant recognition in the visual cortex*. National Resource Laboratory for the Study of Brain and Behavior, University of Rochester, Technical Report 97.2, May 1997.

[Rodrigues, 2002] Rodrigues, F. A. *Localização e Reconhecimento de Placas de Sinalização Utilizando um Mecanismo de Atenção Visual e Redes Neurais Artificiais*. Dissertação de Mestrado, Universidade Federal de Campina Grande, 2002.

[Roelfsema et al., 1998] Roelfsema, P. R.; Lamme, V. A. and Spekreijse, H. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, vol. 395, pp. 376-381, 1998.

[Rosenfeld, 1983] Rosenfeld, A. Image Analysis: problems, progress and prospects. *In: Readings in Computer Vision*. Morgan Kaufmann Publishes, 1983.

[Ross et al., 1974] Ross, M. J.; Shafter, H. L.; Cohen, A.; Freudberg, R. and Manley, H. J. Average magnitude difference function pitch extractor. *In: IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 5, pp. 353-362, 1974.

[Russel and Norvig, 1995] Russel, S. and Norvig, P. *Artificial Intelligence: a modern approach*. Prentice Hall, Englewood Cliffs, New Jersey, 1995.

[Spagnolo et al., 2004] Spagnolo, Paolo; Leo, Marco; D'Orazio, Tiziane and Distante, Arcângelo. Robust Moving Object Segmentation By Background Subtraction. *In: 5th International Workshop on Image Analysis for Multimedia Interactive Services*, Lisboa, April 2004.

[Sun and Fisher, 2003] Sun, Y. and Fisher, R. Object-based Visual Attention for Computer Vision. *Artificial Intelligence*, vol. 146, no. 1, pp. 77-123, 2003.

[Sun and Fisher, 2002] Sun, Y. and Fisher, R. Hierarchical Selectivity for Object-based Visual Attention. *In: Biologically Motivated Computer Vision*, pp. 427-438, 2002.

[Thompson and Schall, 2000] Thompson, K. G. and Schall, J. D. Antecedents and correlates of visual detection and awareness in macaque prefrontal cortex. *Vision Research*, vol. 40, pp. 1523-1538, 2000.

[Treue and Trujillo, 1999] Treue, S. and Trujillo, J. C. M. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, vol. 399, pp. 575-579, 1999.

[Tsotsos et al., 1995] Tsotsos, J. K.; Culhane, S. M.; Wai, W. Y. K.; Lai, Y. H.; Davis, N. and Nuflo, F. Modelling Visual Attention via Selective Tuning. *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, Oct. 1995.

[Wang et al., 2004] Wang, J.; Reinders, M. J. T.; Lagendijk, R. L.; Lindenberg, J. and Kankanhalli, M. S. Video content representation on tiny devices. *In: Proceedings of the IEEE International Conference on Multimedia and Expo*, Taipei, July 2004.

[Wildes, 1998] Wildes, R. P. A measure of motion salience for surveillance applications. *In: Proceedings of the IEEE International Conference on Image Processing*, pp.183-187, 1998.

[Wolfe, 1998] Wolfe, J. W. Visual Search. *In: Attention*, edited by H. Pashler, Psychology Press Ltd., pp. 13-73, 1998.

[Yeasin, 2002] Yeasin, Mohammed. Optical Flow in Log-Mapped Image Plane: a new approach. *In: IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 125-131, Jan. 2002.

[Yee et al., 2001] Yee, H.; Pattanaik, S. and Greenberg, D. P. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *In: ACM Transactions on Graphics*, ACM Press, pp. 39-65, 2001.

Apêndice A

Detecção de Transições Abruptas: Vídeos

Este apêndice apresenta uma amostra dos vídeos utilizados no estudo de caso envolvendo a detecção de transições abruptas (Seção 5.2). Os cinco primeiros vídeos (Figuras 46 a 50) foram utilizados no conjunto de treinamento, enquanto os outros cinco (Figuras 51 a 55) foram utilizados no conjunto de teste.

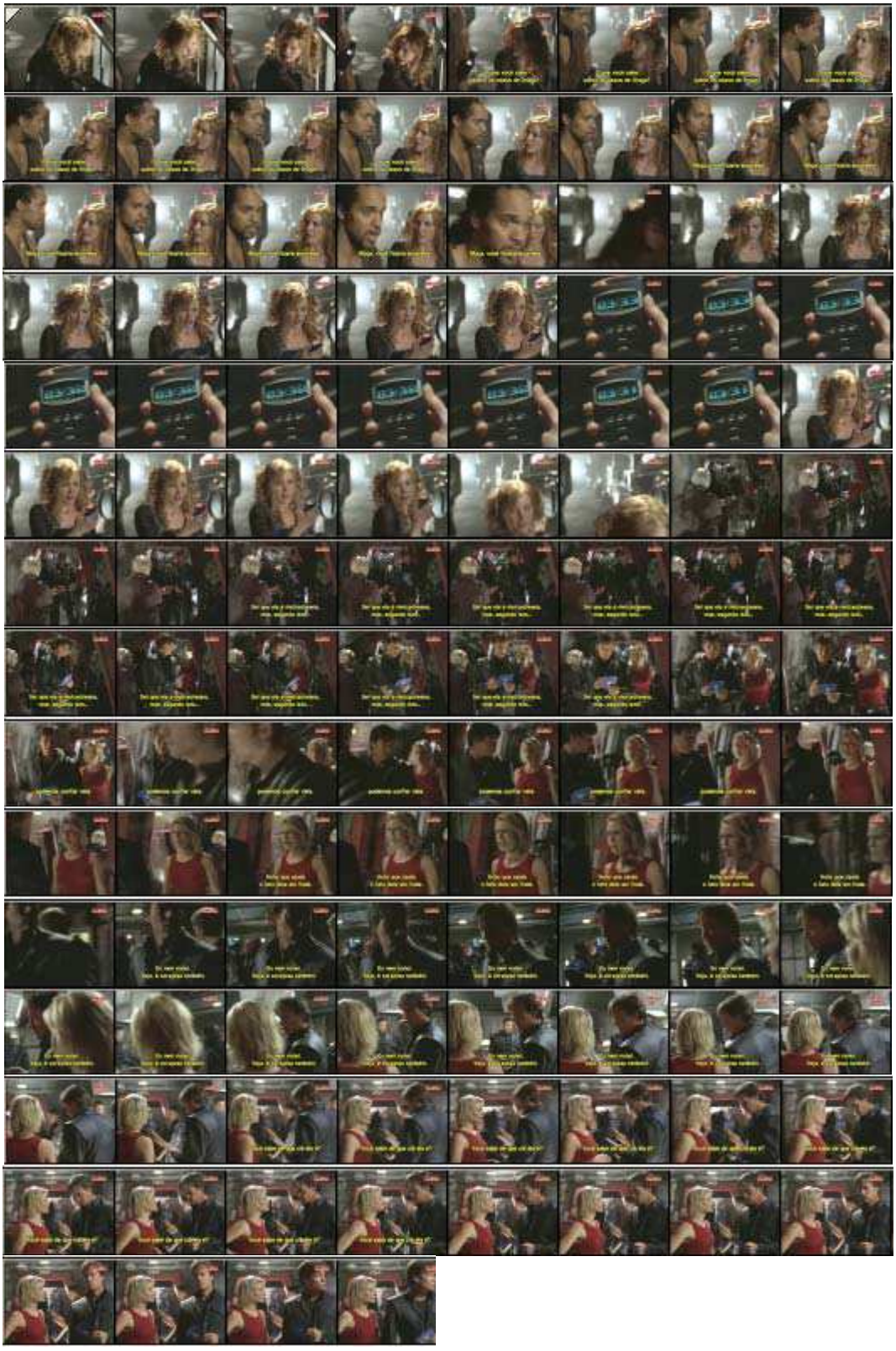


Figura 46 – Amostra de quadros do vídeo 1 do conjunto de treinamento

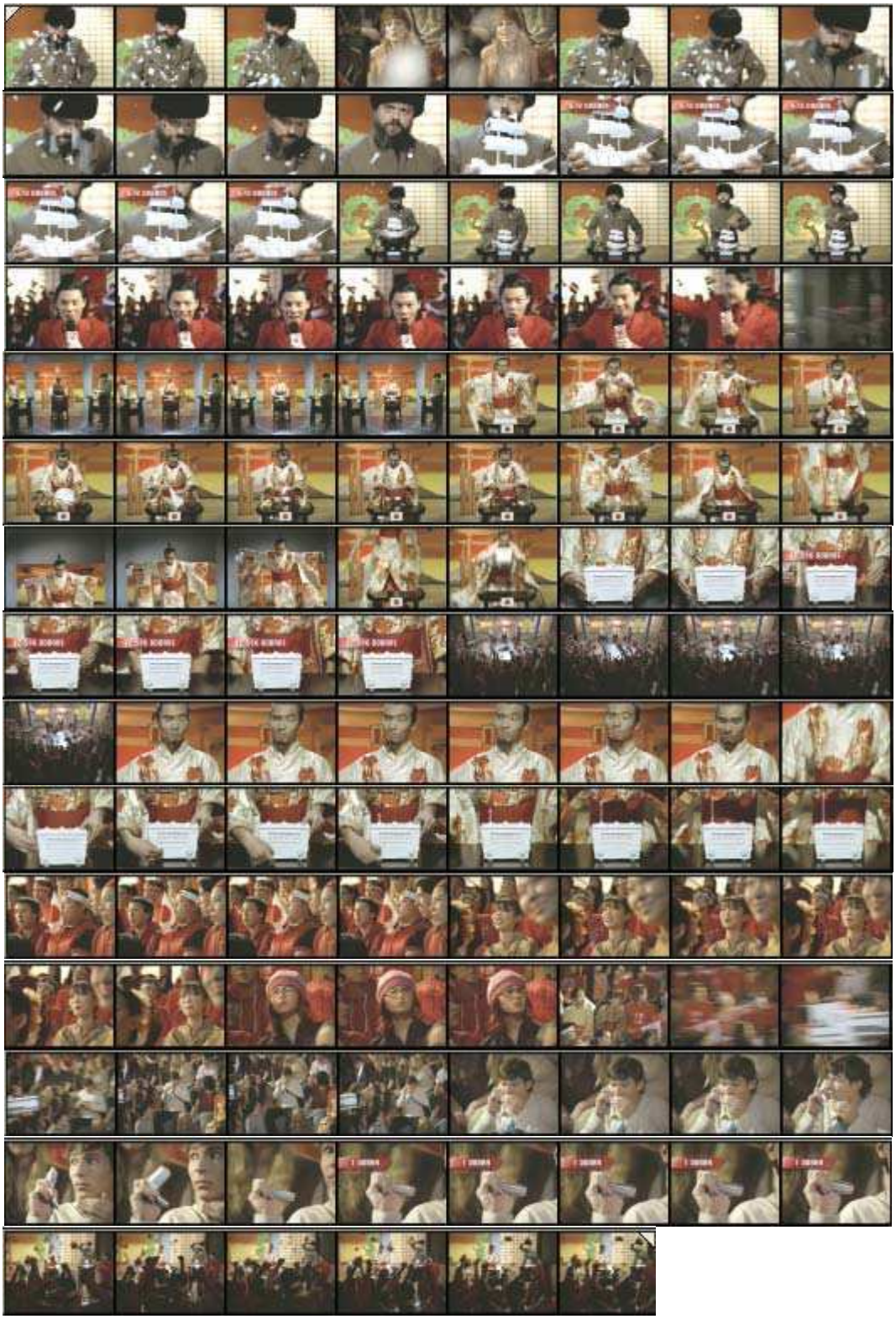


Figura 47 – Amostra de quadros do vídeo 2 do conjunto de treinamento

Figura 48 – Amostra de quadros do vídeo 3 do conjunto de treinamento

Figura 49 – Amostra de quadros do vídeo 4 do conjunto de treinamento

Figura 50 – Amostra de quadros do vídeo 5 do conjunto de treinamento

Figura 51 – Amostra de quadros do vídeo 1 do conjunto de teste

Figura 52 – Amostra de quadros do vídeo 2 do conjunto de teste

Figura 53 – Amostra de quadros do vídeo 3 do conjunto de teste

Figura 54 – Amostra de quadros do vídeo 4 do conjunto de teste

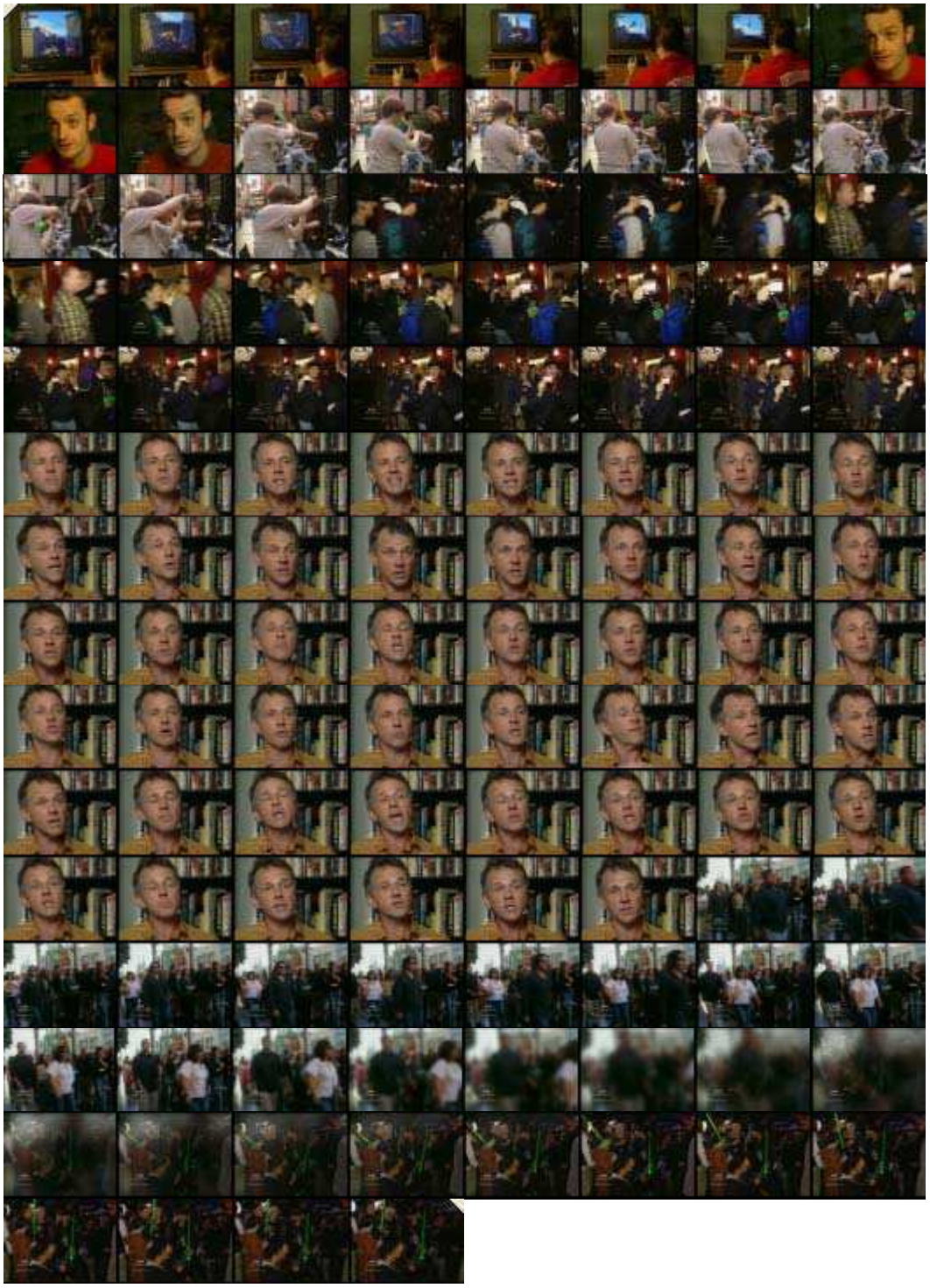


Figura 55 – Amostra de quadros do vídeo 5 do conjunto de teste

Apêndice B

Sistema de Atenção Visual: Vídeos

Neste apêndice, é apresentada uma ilustração dos processamentos realizados pelo sistema de Atenção Visual desenvolvido neste trabalho. Nas Figuras 56 a 59, estão expostas, respectivamente, amostras (i) de um vídeo utilizado nos experimentos; (ii) do vídeo cujos quadros são os mapas de movimento relativos a cada par de quadros originais, gerado pelo Módulo de Atenção Temporal; (iii) do vídeo cujos quadros são os quadros originais segmentados pelo movimento, gerado pelo Módulo de Segmentação; e (iv) do vídeo cujos quadros são os mapas de saliência segmentados referentes aos quadros originais, gerado pelo Módulo de Atenção Espacial *bottom-up*.

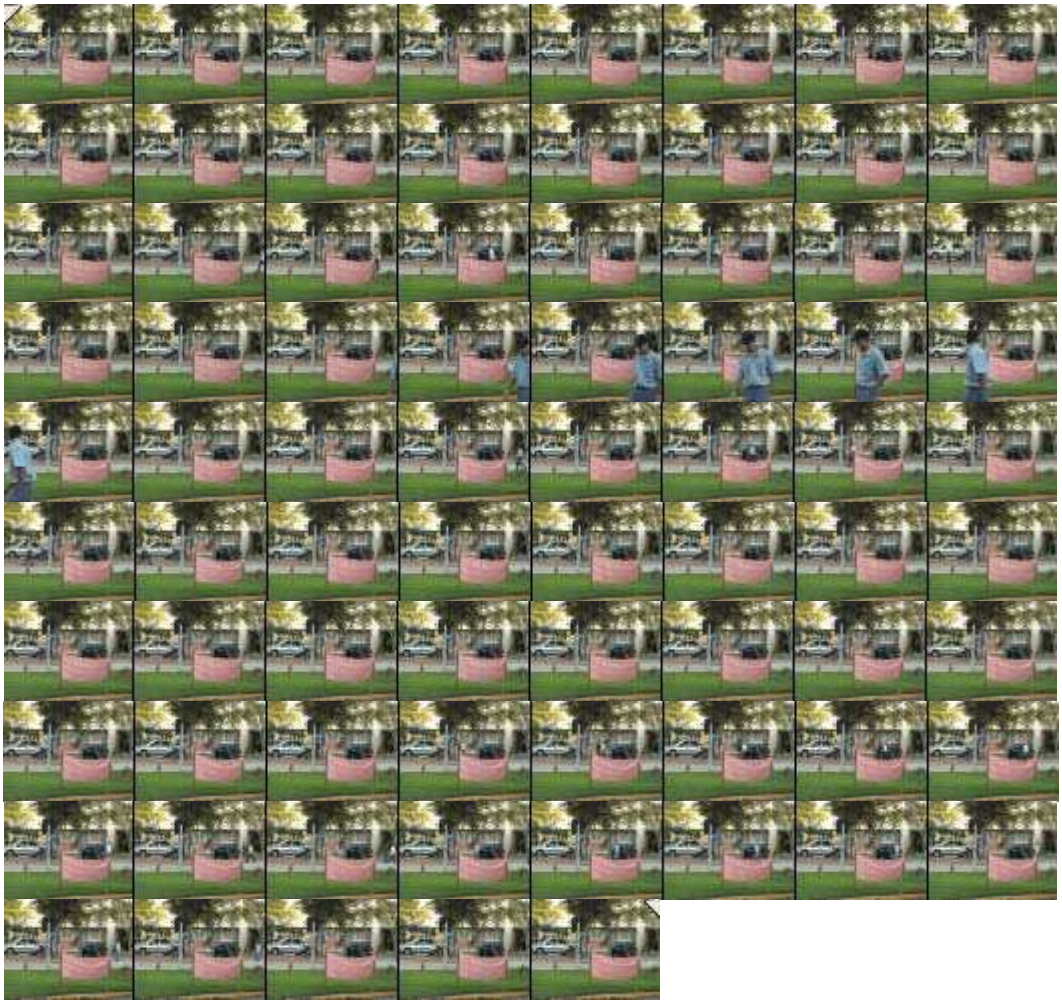


Figura 56 – Amostra do vídeo original

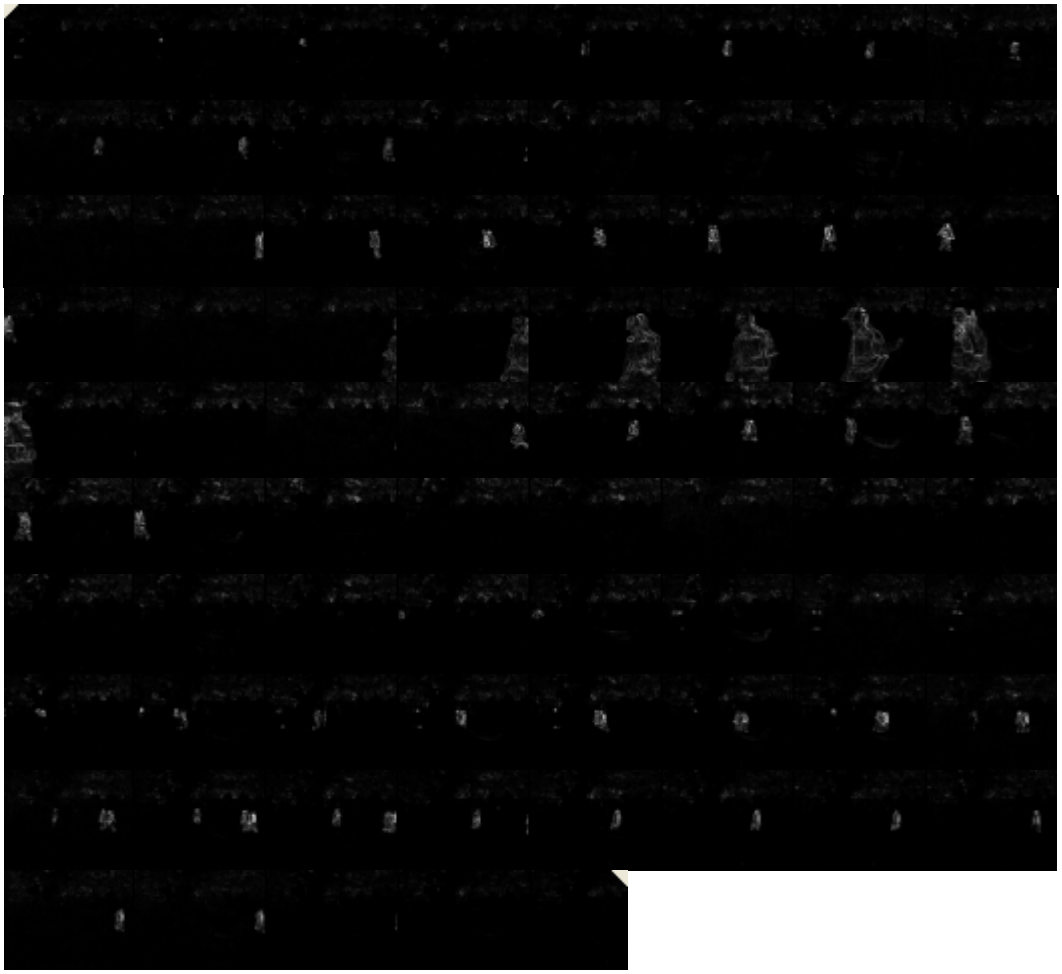


Figura 57 – Amostra do vídeo composto por mapas de movimento

Figura 58 – Amostra do vídeo composto por quadros segmentados

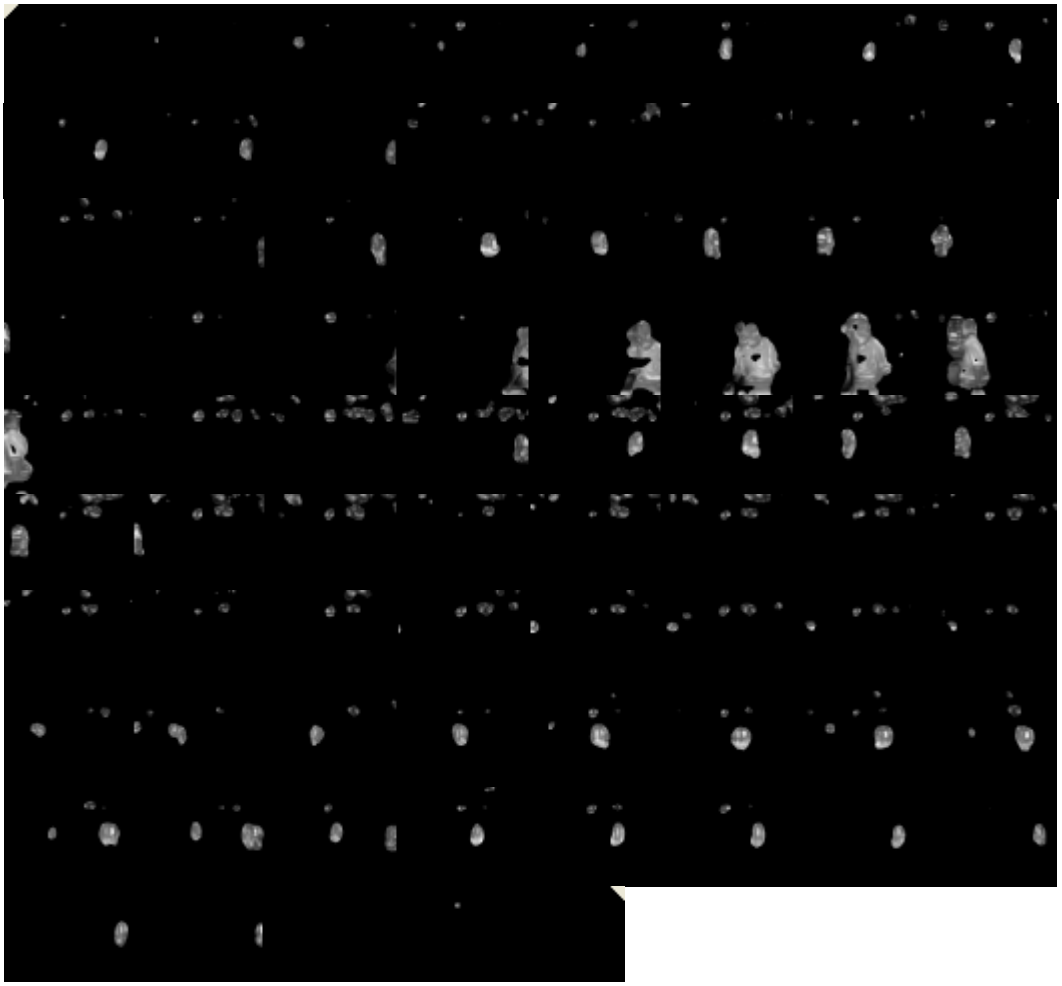


Figura 59 – Amostra do vídeo composto por mapas de saliência segmentados

Apêndice C

Programas e Vídeos

O CD-ROM em anexo contém os códigos-fonte dos programas desenvolvidos neste trabalho, cópias das bibliotecas de C/C++ utilizadas (como, por exemplo, os pacotes *IPP* e *OpenCV*), alguns vídeos utilizados/gerados nos experimentos, cujas amostras foram apresentadas nos Apêndices A e B, e uma cópia eletrônica desta dissertação no formato *pdf*.