Jorge Henrique Busatto Casagrande

DETECÇÃO DE MOVIMENTO ANORMAL EM VIDEOVIGILÂNCIA BASEADA EM RASTREAMENTO E AGRUPAMENTOS UNIFORMES ÓTIMOS

Florianópolis

2015

Jorge Henrique Busatto Casagrande

DETECÇÃO DE MOVIMENTO ANORMAL EM VIDEOVIGILÂNCIA BASEADA EM RASTREAMENTO E AGRUPAMENTOS UNIFORMES ÓTIMOS

Tese submetida ao Programa de Pós Graduação
em Engenharia de Automação e Sistemas
para a obtenção do Grau de Doutor.
Orientador:
Prof. Marcelo Ricardo Stemmer Ph.D.
Universidade Federal de Santa Catarina

Florianópolis

2015

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Casagrande, Jorge Henrique Busatto Casagrande
Detecção de Movimento Anormal em Videovigilância Baseada em Rastreamento e Agrupamentos Uniformes Ótimos / Jorge Henrique Busatto Casagrande Casagrande ; orientador, Marcelo Ricardo Stemmer - Florianópolis, SC, 2015.
174 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia de Automação e Sistemas.

Inclui referências

1. Engenharia de Automação e Sistemas. 2. Detecção de Movimento Anormal. 3. Análise de Vídeo. 4. Reconhecimento de Padrões. 5. Videovigilância Automatizada. I. Ricardo Stemmer, Marcelo. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia de Automação e Sistemas. III. Título.

Jorge Henrique Busatto Casagrande

DETECÇÃO DE MOVIMENTO ANORMAL EM VIDEOVIGILÂNCIA BASEADA EM RASTREAMENTO E AGRUPAMENTOS UNIFORMES ÓTIMOS

Esta Tese foi julgada aprovada para a obtenção do Título de Doutor, e aprovada em sua forma final pelo Programa de Pós Graduação em Engenharia de Automação e Sistemas.

Florianópolis, 20 de fevereiro 2015.

Prof. Rômulo Silva de Oliveira Ph.D.
Coordenador do Programa de Pós-Graduação em Engenharia de Automação e Sistemas - Universidade Federal de Santa Catarina

---

Prof. Marcelo Ricardo Stemmer Ph.D.
Universidade Federal de Santa Catarina
Orientador

Banca Examinadora:

Prof. Adilson Gonzaga Ph.D. - EESC/USP

Prof. William Robson Schwartz Ph.D. - DCC/UFMG

Prof. Rômulo Silva de Oliveira Ph.D. - DAS/UFSC

Prof. Armando Albertazzi Goncalves Junior Ph.D. - EMC/UFSC

Prof. Jomi Fred Hubner Ph.D. - DAS/UFSC

À todas as forças que me fazem forte, inclusive as contrárias! À minha família que motivou, compartilhou e abdicou do que foi possível para permitir minha dedicação a esta obra.

AGRADECIMENTOS

À Universidade Federal de Santa Catarina, referência do que há de melhor no ensino público, a qual me formou Engenheiro, Mestre e agora, Doutor. Ao Instituto Federal Santa Catarina que acreditou e investiu na minha iniciativa de capacitação e me faz sentir orgulhoso e retribuído de ser professor e brasileiro. Ao meu solícito orientador pelos inúmeros encontros e discussões que em meio aos seus conselhos e sugestões encontrei tranquilidade e a certeza de uma grande amizade.

É irreversível a necessidade de se replicar a inteligência humana em máquinas inanimadas. As boas intenções dessa busca contribuem para o bem estar da sociedade, tornando nossa vida mais segura, saudável, confortável e conectada.

(Jorge H. B. Casagrande, 2015)

RESUMO

A videovigilância compõe-se de um conjunto de câmeras e demais recursos tecnológicos para servir como uma ferramenta que visa a segurança pública ou privada em locais estratégicos da movimentação de pessoas e/ou veículos. Os interesses por estes sistemas, em expansão pelo mundo, estão ligados a sua potencialidade em coibir atos antissociais, apoiar na melhoria da mobilidade urbana ou ainda detectar ou prevenir eventos que demandem ação imediata para evitar colapsos, ou mesmo salvar vidas. A automação na monitoração é uma necessidade irreversível pois, sendo centralizada, depende de um operador humano para fiscalizar muitas câmeras através de um trabalho tedioso, cansativo e sujeito a erros e omissões no acompanhamento de movimentação suspeita. A Detecção de Movimento Anormal (DMA) é uma análise de vídeo útil para fins de videovigilância e, em especial, aquela realizada sobre o rastreamento de objetos em trajetos globais não usuais. Em função das barreiras no tratamento computacional de grandes volumes de dados, mesmo nas modernas arquiteturas de sistemas embarcados, propostas encontradas nas abordagens baseadas em rastreamento são geralmente limitadas em flexibilidade no que diz respeito a cenários, metas, duração do vídeo e realidade e assim, nem sempre viáveis nas aplicações em tempo real. Visando extrair o melhor de um modelo estatístico para esse propósito, como o modelo de misturas gaussianas (GMM - *Gaussian Mixture Modeling*), o presente trabalho apresenta uma nova abordagem para DMA ancorada sobre um classificador binário ótimo e construída a partir de três processos iterativos durante um treinamento supervisionado: a geração de amostras sobre agrupamentos uniformes formando uma grade de regiões, a aprendizagem por região dos parâmetros de uma função de distribuição de probabilidade (*pdf*) multivariada e por fim, o uso de curvas características de operação do receptor (ROC - *Receiver Operating Characteristics*) para encontrar o melhor classificador. Como base para avaliar a abordagem foram utilizados dados resultantes de anotações de vídeo do mundo real, elaborados a partir de ferramentas próprias ou de domínio público. Os resultados avaliados demonstraram que cada cenário possui uma área de agrupamento que otimiza o desempenho da DMA mesmo com uma significativa redução de amostras. Neste aspecto, além da tese contribuir com uma metodologia que garante a melhor performance dentro da abordagem de DMA proposta, ela revela que uma análise baseada em região reduz o custo computacional sem afetar significativamente a qualidade das inferências.

Palavras-chave: Detecção de Movimento Anormal, Análise de Vídeo, Videovigilância Automatizada, Análise de Movimento, Reconhecimento de Padrões

ABSTRACT

Video surveillance is composed of a set of cameras and other technological resources to serve as a tool to public or private safety in strategic locations of moving people and/or vehicles. The interest by these systems, expanding worldwide, are linked to their potentiality in curbing antisocial acts, to assist in improving urban mobility or also detect or prevent events that require immediate action to prevent collapses, or even save lives. The automation in monitoring these systems is an irreversible necessity, because being centralized, depends on a human operator to monitor many cameras through a tedious, tiresome and prone to errors and omissions job in the monitoring of suspicious motions. The Abnormal Motion Detection (AMD) is a useful video analysis for video surveillance purposes, and in particular, that performed on the objects tracking in unusual global paths. Due to the barriers in computational treating of large amounts of data, even in modern architectures embedded systems, proposals found in tracking based approaches are generally limited in flexibility regarding scenarios, goals, length of video and reality and thus, not always feasible in real-time applications. Aiming to extract the best from a statistical model for this purpose, as a Gaussian Mixture Model (GMM - *Gaussian Mixture Modeling*), this work presents a new approach to AMD docked on a best binary classifier and built from three iterative processes over a supervised training: The samples generation over uniform clusters forming a grid of regions, learning the parameters per region of a probability distribution function (*pdf*) multivariated and finally, the use of curves the receiver operating characteristics (ROC - *Receiver Operating Characteristics*) to find the best classifier. As a basis to evaluate the approach, data derived from real world video annotations were used, elaborated from own or public domain tools. The evaluated results demonstrated that each scenario has a clustering area that optimizes the AMD performance even with a substantial samples reduction. In this regard, besides the thesis contribute on a methodology that ensures the best AMD approach performance, it reveals that a region-based analysis reduces computational cost without significantly affecting the inferences quality.

Keywords: Abnormal Motion Detection, Video Analysis, Automated Surveillance, Motion Analisys, Pattern Recognition.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LISTA DE SÍMBOLOS

SUMÁRIO

1 INTRODUÇÃO

1.1 A VISÃO COMPUTACIONAL E A VIDEOVIGILÂNCIA

A visão computacional é uma área da engenharia com foco na pesquisa e desenvolvimento de métodos e técnicas para que os sistemas computacionais consigam interpretar imagens em diversos contextos. Diante dos resultados desses estudos, as informações relevantes adquiridas como dados de entrada para inúmeras aplicações na automação de sistemas, deixou de ser tarefa somente de sensores eletrônicos baseados em fenômenos físicos ou químicos como mecânicos, fotoelétricos, térmicos, sonoros, eletromagnéticos, entre outros. As imagens captadas por câmeras passam a substituir com vantagens em diversas aplicações, muitos desses tipos de sensores uma vez que elas tem muitas particularidades com a visão humana. As informações capturadas por essa via carregam um grau maior de detalhes que, se bem modelados, ajudam os sistemas computacionais a tomarem decisões sobre o que estão "vendo", similarmente a reação dos seres humanos quando estão diante de observações dos eventos ou fenômenos de interesse. Isso vale tanto para imagens estáticas quanto para uma sequência delas, quando assim formam uma sequência de imagens do mesmo cenário denominados de *frames* de vídeo.

Baseado nessa ideia, um amplo espectro de aplicações foi sendo somado a um rol de possibilidades que ficou melhor adaptado ou adequado ao campo de estudo da visão computacional. Como consequência, esta área se desdobrou em vários ramos de pesquisa distribuídos entre duas categorias bem definidas de aplicações: aquelas ligadas com o processamento e análise de imagens e as outras associadas com o processamento e análise de vídeo. A Figura 1 ilustra as principais ramificações da visão computacional como uma tentativa de definir sua taxonomia. Os extremos de todos esses ramos geram uma infinidade de tópicos de estudos específicos das várias aplicações que crescem a cada ano, e que por isso não são detalhados nessa classificação.

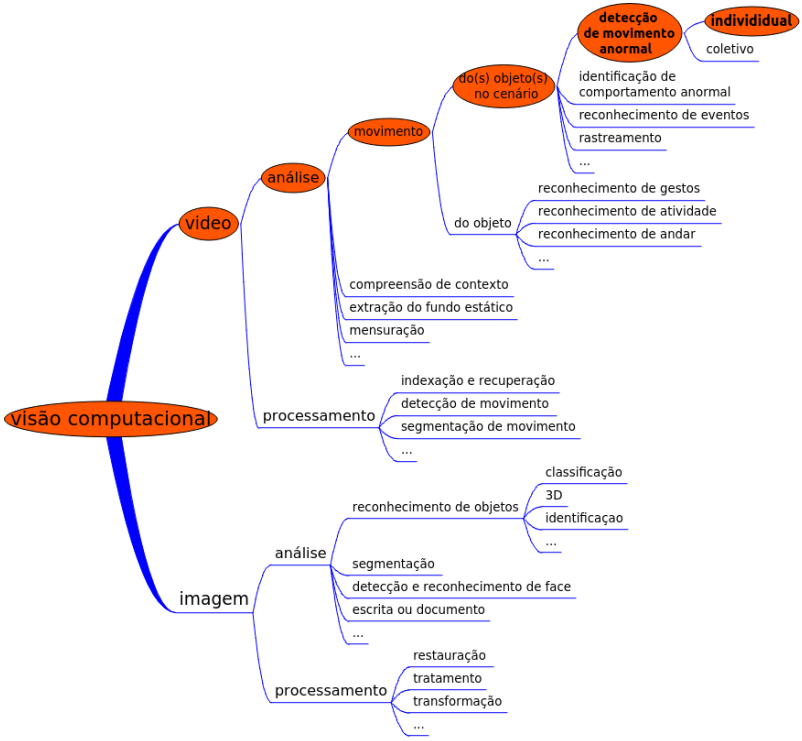
Uma diferenciação importante nessa taxonomia é a separação entre as classes *vídeo* e *imagem*. Em síntese, o vídeo é um conjunto de imagens sequenciadas em uma determinada escala de tempo que proporciona ao observador, a sensação e o registro de movimento dos objetos móveis pertencentes a um mesmo cenário. Toda aplicação que necessita obrigatoriamente usar um ou mais *frames* anteriores como referência de entrada para produzir um resultado de saída, está categorizada como uma aplicação de vídeo. Na divisão dos ramos é importante destacar que o termo *análise* está relacionado com um conjunto de ferramentas mais sofisticado que permite modelar problemas em

um nível maior de abstração. Em geral, os resultados de uma análise estão relacionados com tomadas de decisão ou inferências sobre um conjunto de informações que participou da modelagem. O termo *processamento* refere-se a alguma transformação matemática sobre os dados de entrada que então resultam em outros, agora carregados com novas informações de interesse. Esses novos dados podem ser os dados de entrada para um novo processamento ou análise como os que estão previstos nessa taxonomia.

Desse modo, os tópicos de pesquisa nas aplicações de videovigilância, podem estar situados nas pontas das muitas ramificações da categoria vídeo. Os ramos associados com o trabalho desenvolvido aqui estão circulados com uma elipse na Figura 1. Vale destacar que a expressão *vigilância* está ligada fortemente com o significado do verbo *monitorar* e dessa maneira, a diversidade de aplicações vai além daquelas com o propósito de segurança. Pode-se incluir aí, fins como detecção, identificação, reconhecimento, suporte a automação residencial ou industrial, entre outros. Usando os diversos mecanismos da visão computacional e os modelos propostos ao longo desses últimos anos, especialmente aqueles embarcados com técnicas de inteligência artificial, a análise de vídeo tem explorado tópicos de pesquisa avançados como, videovigilância inteligente, os sistemas interativos de realidade virtual, as interfaces avançadas de percepção humano-computador, a visão bio-inspirada e antecipada, a visão robótica, a indexação e recuperação de vídeo baseado em contexto ou conteúdo, a análise de performances esportivas, a análise meteorológica, os sistemas de suporte ao motorista, os estudos clínicos, a monitoração e controle de tráfego, os sistemas de inteligência ambiente, o suporte a sistemas legados de automação e segurança, entre outros.

Nessa perspectiva, a automação da monitoração de câmeras instaladas em locais estratégicos da movimentação de pessoas e/ou veículos tem estimulado o desenvolvimento de aplicações voltadas à segurança pessoal ou mesmo patrimonial. Encontrar os melhores métodos e abordagens para tornar efetivamente útil essas aplicações é mérito dos ramos de investigação em videovigilância os quais tem recebido uma intensa contribuição nos últimos anos (LI et al., 2012a) (ABRAMS et al., 2012). O reconhecimento da atividade do ser humano, o principal ator ou o objeto nesse contexto, é naturalmente uma das metas mais evidentes. Nesse sentido, a análise da movimentação de objetos dentro de um cenário ou do movimento do próprio objeto é a tarefa de maior nível a ser resolvida nesse escopo. Observa-se que para implementar uma aplicação como essas, é necessário uma cadeia de etapas que envolvem também outros ramos de pesquisa definidos na taxonomia da Figura 1. Por exemplo, para detectar o movimento anormal de um objeto baseando-se em um modelo que acompanha o caminho no tempo e espaço do alvo, é necessário como etapa anterior, modelar um sistema que consiga rastrear o ob-

Figura 1: Os principais ramos de pesquisa em visão computacional.



Fonte: Elaborado pelo autor.

jeto em questão de modo confiável . Essa tarefa anterior vai exigir a aplicação de modelos próprios de domínio de outro ramo da pesquisa em análise de vídeo. Por sua vez, o rastreamento de objetos também vai consumir outros modelos que envolvem domínios da pesquisa como o da segmentação de movimento, situados no ramo de pesquisa de processamento de vídeo.

Ainda, e não menos importante, especialmente no ramo de análise de vídeo, o espectro de tópicos de pesquisa se multiplica caso desejássemos segmentar aqueles trabalhos que consideram se a captação de *frames* é feita por câmera única ou múltiplas, fixas ou móveis, calibradas ou não, entre outras situações. Para manter um foco dentro dessa infinidade de possibilidades, o presente trabalho se apoiou naquelas situações onde há o maior interesse neste ramo de pesquisa. Sendo assim, dirigiu-se a atenção para a videovigilância de sistemas legados com câmera única, fixa e sem calibração.

Propor uma solução para a análise do movimento de objetos, exige uma série de domínios de conhecimentos em torno das diversas e complexas etapas de um sistema como esse. Dentre os diversos domínios, se destacam as teorias, métodos e modelos de reconhecimento de padrões, processamento digital de imagens e da inteligência artificial. A análise inicia a partir da captação ou extração dos sinais de vídeo das câmeras e/ou sensores de diversos tipos direcionados para os cenários de interesse, e continua até a não trivial interpretação sobre a normalidade ou não do comportamento dos objetos em movimento identificados. Para tanto, é necessário reproduzir nas máquinas, as noções de aprendizado e decisão baseadas em uma especialização sobre a massa de dados capturada durante a observação de um cenário alvo.

Muitas alternativas de ferramentas e modelos estão bem consolidados dentro da área da visão computacional para resolver os problemas ou mesmo propor soluções na análise de movimento. No entanto, a abordagem, método ou estratégia faz toda a diferença na qualidade dos resultados e na justificativa de sua utilidade e viabilidade.

1.1.1 A Capacidade Computacional na Automação de Sistemas de Visão

A miniaturização alcançada pela larga escala de integração de semicondutores, possibilitou que sistemas embarcados ficassem junto a complexas e convenientes aplicações voltadas para o ser humano, principalmente aquelas onde os resultados dependem das reações em tempo real do que estão recebendo como dados de entrada. Várias arquiteturas de microcontroladores permitem construir dispositivos completos com hardware dedicado, interfaces de entrada e saída, processadores e memórias de grande capacidade e sistemas operacionais incluindo aplicativos completos.

No entanto, as tecnologias e inovações na extração ou sensoriamento de dados com o propósito de automatizar os modernos processos, não evoluiu assim tão rapidamente quanto aos recursos computacionais. A qualidade ou a confiabilidade desses dados, definidos como informação de entrada, são quesitos tão importantes quanto aos métodos e modelos para seu processamento nos sistemas computacionais. Em síntese, essa etapa prévia contribui inegavelmente para o sucesso e utilidade das aplicações de maneira que a forma de como se coleta dados do ambiente passou a ter uma atenção maior da pesquisa nesses últimos 30 anos e ao que tudo indica, com mais ênfase ainda para os próximos anos.

Nos últimos 50 anos do século passado a ciência dedicou muitos esforços para ampliar capacidades de processamento, armazenagem e transmissão de

dados. De fato as conquistas desses estudos estão materializadas em produtos e serviços que estão disponíveis em máquinas poderosas que hoje cabem na palma de nossa mão. Ao longo do tempo a lei de Moore[1] veio sendo confirmada e o processamento de grandes volumes de dados com uso extensivo de grandes quantidades de memória já não é mais um privilégio de super computadores ou de poucas aplicações nos dias atuais. Segundo Anderson (2009), essa lei deve estar chegando ao fim pois os custos de desenvolvimento e consumo de energia estão cada vez mais altos. Como consequência os esforços dos pesquisadores atuais estão mais direcionados na busca de soluções para otimizar o uso dos recursos de hardware do que propriamente aumentar sua capacidade.

As várias etapas de um completo sistema de visão exigem consideráveis recursos computacionais como memória e poder de processamento para abrir e gerenciar os vários processos em tempo real decorrentes não só das etapas preliminares da aquisição e transformação dos dados (pre-processamento das imagens) como também das etapas intermediárias e finais de análise dos *frames* de vídeo. Apesar das alternativas como a divisão desses processos em sistemas modernos de computação em nuvem ou na virtualização de computadores, é importante destacar que no caso de aplicações voltadas a videovigilância, teremos dezenas ou até milhares de câmeras que estarão gerando um número cada vez maior de *frames* Por Segundo (FPS) para serem processadas. Isso se torna mais crítico ainda se a análise dos *frames* de vídeo depende de longos períodos de observação para produzir suas inferências. Especialmente nas abordagens baseadas em rastreamento para fins de automação da videovigilância, o grande volume de dados e a complexidade dos algoritmos, são considerados relevantes barreiras para o tratamento computacional, mesmo nas mais modernas arquiteturas de sistemas embarcados. Como consequência, muitas abordagens encontradas no presente estudo são geralmente limitadas em flexibilidade no que diz respeito a cenários, metas, duração do vídeo e realidade. Sendo assim, é desejável priorizar propostas dentro da área de visão computacional que minimizem os custos computacionais e complexidades dos seus algoritmos para torná-las então, viáveis nas aplicações em tempo real (NAZARE et al., 2014) (BANG et al., 2012) (JAVED; SHAH, 2008).

A maneira usual de contornar os problemas de sobrecarga computacional nesses casos, é a adoção de modelos onde a análise é feita por agrupamen-

[1] Em 1965 o presidente da Intel, Gordon E. Moore, fez uma previsão sobre o futuro de qualquer dispositivo digital (chip) que é baseado em transistores como componentes principais. Moore estabeleceu que pelo mesmo custo de desenvolvimento, um chip teria um número de transistores dobrado a cada 18 meses. Esta referência acabou sendo uma meta importante para indústria de semicondutores que acelerou muitas pesquisas e processos tecnológicos inovadores e assim proporcionou chips cada vez mais complexos e com custos acessíveis.

tos de *pixels* em cada *frame* de vídeo, dividindo assim, uma tarefa complexa baseada em cada *pixel* da imagem por outras mais leves atuando em uma quantidade menor de regiões de interesse ou ROI (Region of Interest). Neste contexto, dependendo da estratégia utilizada para resolver essas tarefas menores, é necessário definir um tamanho ideal desses agrupamentos de modo que o resultado final se mantenha aceitável dentro da proposta da análise. Encontrar o tamanho ideal de agrupamento passa a ser uma avaliação particular de cada proposta e também uma das discussões desta tese.

1.1.2 Análise de Vídeo em Videovigilância

Em uma aplicação usual, um sistema de videovigilância tem o objetivo de monitorar centralizadamente um conjunto de câmeras de vídeo de diversos tipos, usando uma infraestrutura de rede, comunicação e tecnologias próprias. Essa monitoração abrange uma extensa área de interesse para que nela se proporcione algum nível de segurança, seja pelo registro ou seja pela observação em tempo real das imagens captadas dos objetos alvo. A redução dos custos de hardware específicos para tratamento de vídeo (LI; YIN, 2009) também têm permitido a proliferação de sistemas de videovigilância em ambientes comerciais, industriais e residenciais. Sistemas como esses podem conter centenas de câmeras estrategicamente espalhadas como em aeroportos, áreas urbanas, prédios comerciais, estádios entre outros. Uma solução de infraestrutura completa de centralização dedicada a videovigilância para estações de metrô é abordada por Li et al. (2013). Embora toda a rede planejada e implementada contemple a gerência, controle e armazenamento de dados e alarmes, ela é totalmente dependente da operação humana no que se refere a análise de eventos de vídeo. No entanto os autores demonstram na prática um modelo de rede dedicada a videovigilância que já está preparada para embarcar a inteligência da automação através de algoritmos de análise de vídeo como tantos referenciados ou discutidos no presente trabalho.

No local de convergência concentram-se as informações de vídeo e outras mídias associadas das áreas ou zonas vigiadas. Assim, o gerenciamento e armazenamento das informações ocorrem através do uso de computadores, equipamentos específicos como o DVR (Digital Video Recorder), telas de monitoramento e equipe de vigilância proporcional ao que se propõe monitorar. Mesmo com um conjunto de pessoas proporcional para a tarefa de gerência, existem dificuldades para o ser humano tratar e reconhecer todos os eventos importantes.

O volume de informações, mesmo sendo resultado do uso das melhores técnicas de compressão de vídeo, atinge grandezas que são difíceis de

tratar e transmitir, especialmente na condição de tempo real. Considerando que essas informações convergem para uma central de monitoração, ainda é necessário levar em conta que as infraestruturas necessárias de comunicação e conectividade possuam larguras de banda[2] suficientes para viabilizar este tipo de serviço (YE et al., 2013).

Uma solução para aliviar os gargalos de transmissão, armazenamento e tratamento de informação dos sistemas de videovigilância é o uso da análise de vídeo automatizada como chave para que o sistema determine quando começar e quando terminar o uso dos recursos da infraestrutura para tratar eventos efetivamente relevantes. Destaca-se nesse ponto que, como elemento fundamental, o operador humano também faz parte desse sistema e ele possui severas limitações na atividade de monitoração (MORRIS; TRIVEDI, 2008). Grande parte da fiscalização visual ainda depende de um operador humano na filtragem do grande volume de informações de vídeo. É um trabalho tedioso, cansativo e sujeito a erros e omissões no acompanhamento de eventos relevantes que em geral, ocorrem de forma rara e aleatória.

Os sistemas com pouca ou nenhuma inteligência só são úteis como ajuda valiosa no resgate de informações acerca dos atos antissociais já consumados uma vez que não possuem a capacidade de gerar alarmes em tempo real. Esse é um fator limitante da eficiência dos sistemas de videovigilância convencional. Os dados de vídeo atualmente são utilizados apenas como uma ferramenta de retrospectiva forense, perdendo assim o seu principal benefício desejado como um sistema ativo em tempo real. Segundo Xu et al. (2010) e Nazare et al. (2014) os sistemas convencionais analógicos conhecidos como circuitos fechado de Televisão (CCTV - Closed-Circuit Television) conectados à DVRs e monitores são a base dos sistemas de primeira geração de videovigilância. Na sequência, a segunda geração de sistemas são aqueles que incorporaram câmeras digitais e sistemas computacionais que permitem dar suporte aos operadores ou mesmo automatizar muitas tarefas através de algoritmos de detecção, rastreamento e classificação visando o reconhecimento de comportamento ou anormalidades. A terceira geração usa câmeras distribuídas e heterogêneas além de sensores de diversos tipos para abranger grandes áreas de vigilância ampliando o uso no suporte a segurança pública e reforço na proteção contra ataques terroristas. Na atualidade estamos vivendo a quarta geração de sistemas de videovigilância onde sistemas são propostos para fornecer alertas de eventos em tempo real e padrões estatísticos de longo prazo e em sistemas vídeo vigilância distribuída em larga escala. Exemplo

[2] A largura de banda em redes e meios de comunicação está diretamente ligada com o desempenho na transmissão e encaminhamento das informações dos sinais relacionados com os *frames* de vídeo capturados pelas câmeras. Ela pode ser definida como o intervalo de frequências contido no sinal ou que um canal de comunicação permite passar ou ainda definida como a quantidade de bits por segundo que um canal ou uma rede é capaz de transmitir.

disso é o pioneiro sistema S3 (Smart Surveillance System) da IBM Corporation (HAMPAPUR et al., 2007) que foi construído sob infraestrutura de rede IP (Internet Protocol) usando infraestruturas variadas de redes, sistemas distribuídos, base de dados e câmeras IP. O sistema além de monitorar automaticamente a cena através de aplicativos instalados como *plugins*, desempenha o gerenciamento de dados, recupera vídeos indexados baseados em diversos reconhecimentos de eventos e reporta alarmes em tempo real via web.

Neste contexto, a tecnologia de detecção e monitoração de comportamento de objetos móveis nos diversos ambientes requer conhecimentos científicos e tecnológicos em diversas áreas em visão computacional, reconhecimento de padrões e de comportamento, análise de redes de sensores sem fio, tratamento eficiente de fluxo de vídeo, entre outros. Algumas dessas teorias têm sido consolidadas e transferidas em produtos comerciais como, sistemas oferecidos pela Panasonic, Bosch, Siemens, Vidient e Vistascape (LI; YIN, 2009). Esses sistemas possuem muitas funções úteis como detecção de mudança de direção, contador de alvos em movimento passando por uma região, detecção de um intruso em uma área proibida, objetos esquecidos entre outras (TITTA et al., 2011) (JAVED; SHAH, 2008). Eles prometem a garantia de identificar dentro de determinados contextos, mudanças das atividades do ser humano que é o principal ator observado nas zonas monitoradas.

O estudo feito por Räty (2010) mostra que a automatização de sistemas que envolvem a interpretação das cenas de vídeo não só tem um forte apelo social como também uma aplicabilidade muita ampla em vários setores da sociedade como engenharia e medicina. Os entraves tecnológicos destas aplicações estão aos poucos deixando de ser importantes pois estes recursos estão cada vez mais acessíveis. A criação de bases de dados para futuras indexações automáticas de vídeo em todos os níveis de processamento das imagens também já é outro recurso comum. Os recursos tecnológicos não estão somente ligados ao aumento da capacidade computacional e da resolução e qualidade das câmeras atuais. Inclui-se aí o processamento em multi-núcleos de GPU (Graphics Processing Unit) (ZHANG et al., 2012) que motiva o uso de imagens 3D (três dimensões), aos protocolos de controle, redes IP (Internet Protocol) e Câmeras IP incluindo o uso em sistemas embarcados (APPIAH et al., 2009) (YANG et al., 2009).

Os trabalhos de Ma e Wan (2009) e Xu et al. (2010) destacam como a associação de dados originados em outros sensores que não somente câmeras, pode ampliar a confiabilidade no rastreamento de objetos em diversos cenários. Para este fim pode-se utilizar a fusão de outras diversas fontes de dados como áudio, identificação por etiquetas ativas ou passivas de RFID (Radio Frequency IDentification), por sinal de rádio ou geo-coordenadas (GPS - Global Positioning System), satélites além do uso de múltiplas câmeras incluindo

redes de sensores (CHEN et al., 2008). Essa diversidade de opções na busca de contribuições para resolver os mais diversos problemas em videovigilância tem apresentado importantes progressos, indicando que há um campo vasto de pesquisa nesta área.

O vídeo-monitoramento automatizado através de múltiplas câmeras por exemplo, é especialmente importante ao grande número de áreas sensíveis à segurança tais como bancos, lojas, residencias, estacionamentos, etc, seja em tempo real ou através das sequências de vídeo armazenados em arquivos eletrônicos ou gravados em mídias apropriadas. Em ambientes que já possuem algum sistema de segurança baseado na monitoração de circuitos internos ou externos de câmeras, os fluxos de vídeo desses arquivos são usados somente após um evento de interesse ter ocorrido e de forma simples, como uma ferramenta de identificação de detalhes dos eventos de interesse.

Essas aplicações trazem uma boa carga de motivos para utiliza-las como inspiração e base na construção do presente trabalho. Na vida real, implementações como essas podem evitar ou coibir atos criminosos, terroristas, prejuízos financeiros ou a própria vida humana ou ainda simplesmente identificar movimentos não usuais nas cenas de vídeo que precisam receber uma atenção maior de um equipe de segurança (REVATHI; KUMAR, 2012). Direcionando a videovigilância automatizada para o controle de tráfego por exemplo, nas áreas urbanas ou em rodovias, é possível ajudar na melhora da mobilidade ou ainda detectar ou antecipar eventos relevantes que necessitem de ação imediata para evitar colapsos, congestionamentos ou mesmo salvar vidas.

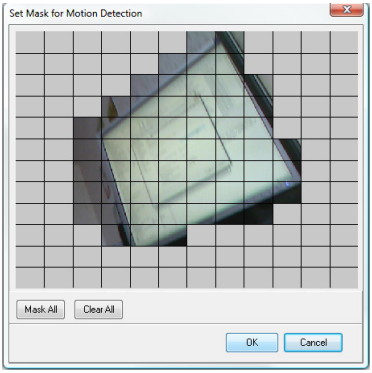
1.1.3 A Análise de Movimento

Um típico exemplo de automação na análise de vídeo é a implementação da detecção de movimento na sua concepção mais simples ou seja, definem-se sub regiões da cena onde se deseja que a gravação dos *frames* de vídeo só seja disparada se o nível médio de intensidade dos *pixels* de qualquer das sub regiões ou soma delas atinge determinado valor. Esse valor é manualmente configurado como um limiar de sensibilidade. Essa é uma ferramenta útil e facilmente configurável que pode ser embarcada nas próprias câmeras, softwares específicos ou nos DVRs. A Figura 2 mostra dois exemplos comuns desse tipo de configuração usando uma máscara disposta sobre a cena através de softwares desenvolvidos exclusivamente para sistemas de videovigilância. A Figura 2(a) mostra uma tela do aplicativo da Webcam Surveyor[3] onde as sub regiões de interesse de monitoração são selecionadas manualmente, já na

[3]Disponível em: http://www.webcamsurveyor.com. Acesso em jul. 2014.

Figura 2(b) é mostrada uma tela de um aplicativo de uso livre para ambientes Linux chamado ZoneMinder[4] onde a forma da área de interesse é livremente escolhida, incluindo limiares de detecção diferenciados.

Figura 2: Exemplos da disposição de máscaras na detecção de movimento em aplicações reais.



(a) Máscara formada por sub-regiões de uma grade uniforme de áreas.

(b) Usando marcações livres de áreas na cena.

Fonte: Extraídos dos sites da Webcamsurveyor e Zoneminder.

Embora esse recurso seja útil e minimize fortemente os gargalos computacionais apontados na subseção anterior, ele atua somente de forma programada para alertar regiões de acesso proibido ou reduzir alarmes falsos ou gravações de *frames* por consequência de movimentação sem interesse no cenário, como por exemplo as folhas de uma árvore devido ao vento. Dependendo do cenário, nem sempre é possível encontrar um limiar de sensibilidade ideal para evitar alarmes ou gravações desnecessárias. Na realidade este tipo de detector foi concebido para ser acionado somente pela mudança de região de algum objeto dentro da área de interesse ou pela presença de algum objeto que entra na área programada para a detecção. Nesse caso estamos tratando na verdade, de um detector de presença mais elaborado que usa conjuntos de *pixels* dos *frames* de vídeo como os sensores do detector. Se disponível e bem programado para gerar alarmes sonoros ou visuais em uma tela de monitoração, este tipo de detector pode ajudar a chamar a atenção do operador para os eventos pré configurados como proibidos somente a partir daquele momento do alarme. Como exemplo de detectores de movimento mais elaborados é o resultado do trabalho de Fang et al. (2009). Os autores abordam uma

[4]Disponível em:http://www.zoneminder.com. Acesso em jul. 2014.

proposta que é focalizada na criação de uma rica interface gráfica para ajustes mais detalhados e adequados para o monitoramento do método proposto. Já o trabalho de Ezzahout e Thami (2013) se concentra na análise do comportamento de pessoas em longos períodos de duração em regiões previamente estabelecidas no aplicativo especificamente elaborado para este fim.

O detector discutido acima é útil para alguns casos, mas a maioria de cenários de videovigilância requer muito mais do que a simples detecção de movimentação de objetos em áreas proibidas pois os cenários reais normalmente são mais complexos, variados, públicos e com muita movimentação de vários tipos de objetos como pessoas, veículos, animais, vegetação e até fenômenos climáticos (chuva, neblina, etc.). O nível de interpretação do movimento precisa ser mais elaborado para ser efetivamente útil como ferramenta de apoio à um operador que precisa monitorar muitas câmeras (BANG et al., 2012).

O passo seguinte na detecção de movimento é o desafio de classificá-lo como um movimento anormal, ou seja, atípico do que tem ocorrido normalmente na observação de muitas horas de monitoração. Isso exige rastrear e armazenar todo o caminho traçado para cada tipo de objeto durante sua passagem na área de interesse da cena. Essa tarefa reúne uma enorme quantidade de informação de entrada que precisa necessariamente ser processada com algum modelo estatístico adequado para classificar anormalidades nos padrões de dados a medida que eles estão sendo recebidos.

Então, uma automação ideal deveria realizar a tarefa de Detecção de Movimento Anormal (DMA) gerando também alertas sonoros e/ou visuais para o operador, dando a ele um efetivo suporte, tanto no acompanhamento *online* quanto em *offline* mediante marcações indexadas em vídeo.

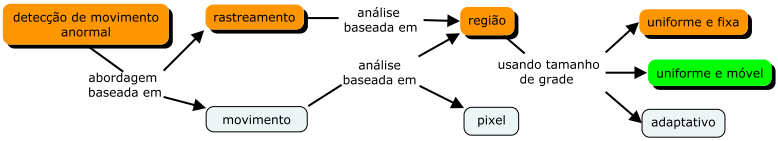
Identificar comportamentos suspeitos ou detectar movimentos anormais também tem uma importância nos sistemas de videovigilância no sentido em que os processos de transmissão e/ou armazenamento das imagens somente devem iniciar quando eventos não desejados são detectados durante o processo de análise de vídeo. Do contrário, todo o volume de informações de vídeo considerados normais acabam utilizando uma energia desnecessária e até impedindo a atenção a eventos realmente relevantes no propósito da monitoração.

Os desafios para automatizar esses processos usando um nível maior de interpretação do movimento continuam sendo alvo de pesquisa e o presente trabalho pretende contribuir nessa direção. Em linhas gerais, espera-se que uma implementação com esse propósito seja confiável na detecção de anormalidades mantendo um baixo índice de falso-positivos além de apresentar um reduzido consumo computacional. Em implementações reais é oportuno embarcar essa automação junto ao próprio hardware de câmeras mas não es-

quecendo de considerar seus recursos computacionais limitados. Esses requisitos foram os principais pontos de partida usados neste trabalho.

A Figura 3 mostra um mapa conceitual que segmenta os trabalhos de vários autores em abordagens bem definidas na pesquisa em análise de movimento anormal. O presente trabalho se posiciona nos ramos que estão destacados com sombreamento e em outra cor. Além de apresentar um método novo na detecção de anormalidades em vídeo, comparando com os resultados com alguns dos principais trabalhos nesta linha, este trabalho também propõe uma solução nova que usa tamanho de grade uniforme e móvel.

Figura 3: Mapa conceitual na segmentação de trabalhos voltados à DMA.



Fonte: Elaborado pelo autor.

As análises baseadas em região, em geral, usam heurísticas ou soluções mais elaboradas para definir o tamanho ideal da região o qual melhor se adapta ao respectivo estudo de caso (ERMIS et al., 2008; KIRYATI et al., 2008; SHI et al., 2010; HANAPIAH et al., 2010; LI et al., 2012a; FEIZI et al., 2012; HAQUE; MURSHED, 2012; CONG et al., 2013). No entanto, vários desses autores usam uma abordagem baseada em movimento para inferir sobre o movimento anormal em uma quantidade pequena de *frames*. Esse segmento de trabalhos é atraente porque não exige qualquer pré-processamento de vídeo ou *framework* de processos que se inicia desde a captação de vídeo. As propostas baseadas em movimento, ficam portanto mais comuns e úteis nos tópicos de pesquisa como análise de andar do objeto, de gestos, de expressões do rosto, do movimento denso de objetos (multidão, tráfego, etc), entre outros.

Embora muitas estratégias baseadas em movimento sejam também úteis na DMA, há limitações de sua aplicabilidade visto que com poucos *frames* de vídeo nem sempre é possível detectar situações anormais como aquelas onde se deseja avaliar trajetos completos dos objetos. Isso leva a uma condição de cenário controlado, diferente da realidade colocada pela videovigilância real. Um sistema efetivamente automatizado deveria considerar vídeos com longo tempo de duração de modo que durante a fase de treinamento sejam acumulados um número elevado de dados para inferir sobre o comportamento usual e individual dos diversos tipos de objetos participantes do cenário observado. Nesse aspecto, os modelos baseados em rastreamento

são mais adequados.

1.2 IDENTIFICAÇÃO E DESCRIÇÃO DO PROBLEMA

A presente tese é desenvolvida com foco no problema da qualidade da avaliação e tomada de decisão sobre anormalidades do movimento de múltiplos objetos rastreados especialmente por longos períodos de observação, típicos da videovigilância real.

A análise de movimento é o nível mais alto de abstração de um sistema e ela depende diretamente da precisão e quantidade dos dados gerados em vários processos anteriores, que por sua vez já consomem relativo esforço computacional. Para concluir sobre anomalias dos movimentos de objetos, ou seja, dos seus trajetos, é necessário preferencialmente um rastreamento confiável e uma eficiente técnica na extração de descritores que represente, com algum grau de fidelidade, todos os objetos alvos.

Os modelos estatísticos continuam ocupando espaços na pesquisa como ferramentas na solução de problemas que emergem como desafios nas várias etapas de análise de vídeo, desde o pré-processamento dos *frames*. No entanto, a grande quantidade de dados necessária para ser processada e a complexidade dos algoritmos, se apresentam como uma barreira ou até mesmo um impedimento na implementação desses modelos em aplicações do mundo real.

Uma estratégia usual para a detecção de movimentos anormais é o uso de classificadores binários onde um limiar de decisão é a referência para definir se um valor calculado faz parte ou não de uma classe. Em modelos estatísticos como o GMM, esse valor limiar pode ser definido empiricamente ou analiticamente através do cálculo de uma probabilidade por técnicas para este fim. O problema então é definir qual o valor de limiar que melhor representa a discriminação e consequentemente a melhor qualidade da decisão. Nesse sentido, observou-se que as características conceituais do uso de curvas ROC são apropriadas para ajudar nessa tarefa e desse modo elas foram utilizadas como ferramenta fundamental na metodologia desenvolvida neste trabalho.

Em adição, na direção de aliviar a grande quantidade de dados gerados no modelo estatístico adotado e consequente custo computacional, a metodologia utilizada foi elaborada sob a luz de uma abordagem de análise baseada em região. Trabalhos de outros autores tem utilizado a análise de movimento em regiões maiores que um *pixel*, subdividindo a ROI, em agrupamentos uniformes de *pixels*. Nessas abordagens predomina o uso de heurísticas ou em alguns casos, soluções mais complexas para definir o tamanho ideal da região que melhor se adapta aos seus casos de estudo. Portanto, o tamanho ideal

deste agrupamento é um outro ponto não devidamente ou claramente explorado nas abordagens de análise de movimento baseadas em região. Neste contexto, o presente trabalho também incorporou no método, a capacidade de encontrar o tamanho ideal de agrupamento de *pixels* para cada cenário vídeo monitorado e assim, contribuindo como um referencial plausível para aqueles trabalhos na mesma linha que necessitam de algum critério para tal tarefa.

1.3 OBJETIVO GERAL

Esta tese apresenta uma abordagem nova para detectar movimentos anormais em cenas reais de videovigilância levando em consideração o problema da carga computacional exigida no reconhecimento de padrões na análise de movimento, especificamente quando se adotam modelos estatísticos na sua concepção. A partir de cenários vídeo monitorados por longos períodos com câmera fixa, os resultados alcançados pretendem ser úteis na modelagem ou implementação de sistemas de videovigilância autônomos ou automatizados dotados com a função de detectar em tempo real, anormalidades de trajetos.

Nesse propósito, o presente trabalho implementa e discute modelos e metodologias construídos a partir de uma abordagem baseada em rastreamento dos objetos móveis que atravessam o campo de visão (Field Of Vision - FOV) de um sistema de monitoração de imagens. Sendo assim, detectar anormalidades de trajetórias globais de objetos móveis é o centro da atenção na análise de vídeo implementada aqui, independente do contexto onde eles participam.

1.4 ESTRUTURA DO TRABALHO

A partir da contextualização explorada neste capítulo a qual revela potenciais eixos de pesquisa nessa área, elaborou-se esse documento. Aqui será relatada uma síntese sobre os principais resultados alcançados nas abordagens conexas pesquisadas até o momento, bem como todos os caminhos explorados para atingir o objetivo geral da tese, dividindo o conteúdo da seguinte forma:

No capítulo 2 discute-se sobre as motivações e os principais eixos da pesquisa neste tema uma vez que é crescente e dinâmico o número dos trabalhos dentro do campo da área de visão computacional que apresentam alternativas de solução na análise de movimento anormal em vídeo voltadas para aplicação no mundo real. Este capítulo também apresenta um levantamento de trabalhos relacionados sobre este tema, destacando as abordagens usuais

desde o modelamento até as técnicas bem consolidados na análise de comportamento de objetos móveis baseados em seus diversos descritores. Em geral a discussão coloca o ser humano como centro das atenções visto que na maior parte das aplicações ele é o ator mais significativo nos ambientes de vídeo monitoração, além de identificar alguns desafios que continuam abertos nestas aplicações. Tanto neste quanto para os próximos capítulos, as análises do ferramental envolvido em um tema tão complexo como esse estarão limitados as suas fundamentações sem o propósito de explorar o enorme arcabouço de matemática e estatística associados.

A tese é explorada no capítulo 3. Nele serão abordados os pontos centrais desta pesquisa onde se destacam a estratégia do uso de grades de regiões visando reduzir o custo computacional através da redução da massa de dados envolvidas na análise de vídeo e o uso das curvas ROC para encontrar limiares ótimos de decisão em classificadores binários construídos com dados de modelos estatísticos como o GMM. Descreve-se também algumas propostas para avançar na direção de tornar viável e confiável o uso da abordagem em sistemas em tempo real graças a redução da dimensionalidade dos dados. A implementação e detalhamento da modelagem da cena, do movimento e da aprendizagem são apresentados sob a luz do estado da arte na detecção de movimentos anormais. As premissas e restrições bem como os pontos positivos e negativos da proposta são discutidos para consolidar a tese com contribuições relevantes no tema.

Os principais resultados são demonstrados no capítulo 4 sob a ótica da base de dados de *datasets* selecionados e adaptados para o propósito da tese. As simulações mais representativas são apresentadas e discutidas.

O capítulo 5 avança um pouco mais sobre as potencialidades da abordagem visando contribuir para os desafios da generalização buscados em reconhecimento de padrões aplicados ao mundo real. Da mesma forma se discute como é possível a aplicação da abordagem na análise de movimento entre câmeras (mosaico de monitoração).

Finalmente no capítulo 6, são descritas as considerações finais sobre os resultados alcançados e as contribuições identificadas do presente trabalho, apresentando ainda uma série de sugestões para futuros trabalhos ancoradas com as novidades encontradas aqui.

2 DETECÇÃO DE MOVIMENTO ANORMAL EM VIDEOVIGILÂNCIA

Os autores Patrick e Bourbakis (2009) destacaram o crescimento substancial da atenção dada pelos pesquisadores sobre a automatização dos sistemas de videovigilância nos 10 anos que antecederam a publicação deles. Desde então as experimentações e propostas tem utilizado métodos de rastreamento de objetos dentro do campo de visão de uma única câmera. Essa linha de pesquisa também tem servido como base para outros ramos de abordagens que envolvem a análise das relações de movimentos de objetos em múltiplas câmeras, incluindo aí as móveis.

As contribuições iniciais para análise de vídeo dirigiram métodos específicos voltados para a vigilância de tráfego em ambientes ao ar livre (externos ou *outdoor*) considerando muitas das variáveis que exigem tratamento especial em função de fatores climáticos tais como vento, precipitações e iluminação natural ou artificial nas suas várias formas de manifestação e mudanças. Por outro lado, o interesse na automatização da vigilância em ambientes internos (ou *indoor*) também tem crescido diante da prevalência dos sistemas de vigilância existentes (sistemas legados) em edificações públicas e privadas e da própria explosão do mercado de equipamentos para essa área. As tecnologias de câmeras CCD (Charge-Coupled Device) e CMOS (Complementary Metal-Oxide Semiconductor), câmeras térmicas e dispositivos de visão noturna são os três dispositivos mais utilizados no mercado de videovigilância (REVATHI; KUMAR, 2012).

Artigos publicados em muitas conferências, workshops e simpósios internacionais do Institute of Electrical and Electronics Engineers (IEEE) exploram as atividades de pesquisa em diversos tópicos de visão computacional. Dentre os vários, pode-se citar alguns que trazem muitas contribuições em análise de movimento e videovigilância como: Conference on Advanced Signal and Video Based surveillance (AVSS); European Conference on Computer Vision (ECCV); International Conference on Pattern Recognition (ICPR);Computer Vision and Pattern Recognition (CVPR); Winter Conference on Applications of Computer Vision (WACV) e International Conference on Computer Vision (ICCV).

Outros eventos internacionais não patrocinados ou organizados pelo IEEE, tem se destacado devido a qualidade dos trabalhos, dos comitês organizadores e do elenco de revisores como o International Conference on Pattern Recognition Applications and Methods (ICPRAM), International Conference on Computer Vision Theory and Applications (VISAPP) e Conference on Graphics, Patterns and Images (SIBGRAPI).

Os laboratórios e grupos de pesquisa, grande parte deles fortemente apoiados pela iniciativa privada, estão investindo em inúmeras aplicações que vão além do escopo usual dos campos de estudo das engenharias e da computação. Eles estão criando inovadores centros científicos e especialidades responsáveis por uma nova geração de pesquisadores.

A pesquisa sobre a identificação ou detecção de comportamento de movimento tem levado autores a exercitarem a criatividade na busca de abordagens e algoritmos mais robustos, leves, confiáveis e insensíveis às variáveis externas. Esses quesitos influenciam demasiadamente na determinação de restrições da implementação, tanto do próprio algoritmo quanto nas próprias aplicações. As variáveis externas estão associadas na captura e pré-processamento de vídeo e são oriundas das distorções da captura dos *pixels* das câmeras, na falta de calibração desses sensores, nas influências da variação da iluminação, na presença de obstáculos e oclusão, na interpretação do fundo estático da cena (*background*), na falta ou excesso de resolução, na segmentação do movimento, entre outras.

Os assuntos apresentados neste capítulo são uma síntese dos tipos de ferramentas e abordagens mais citadas no estado da arte na detecção de anormalidades de movimento em videovigilância e os problemas mais comuns que continuam sendo alvos de pesquisa.

2.1 AS TÉCNICAS DE RECONHECIMENTO DE PADRÕES COMO BASE NA DETECÇÃO DE ANORMALIDADES

A monitoração automatizada dos ambientes está intimamente ligada com o reconhecimento de padrões de comportamento. Esses comportamentos normalmente são aleatórios e previsíveis enquanto estão ocorrendo e portanto possuem uma natureza estatística. Em razão disso, é desejável um modelo que reproduza os estados prováveis do rastreamento dos objetos móveis no campo visual.

As incursões de modelagem e prototipação da abordagem proposta neste trabalho se apoiou na meta de encontrar arranjos ótimos de amostras apoiados no uso de treinamento baseado em reconhecimento, ou seja, aqueles que usam a taxa de erro aparente Bishop (2006) como métrica para avaliar o desempenho de um classificador. De qualquer modo, mesmo não sendo a meta do presente trabalho, procurou-se por um classificador que também pudesse levar o modelo de aprendizagem para um certo grau de generalização na detecção de anormalidades, de modo que seja possível sua utilização em um espectro maior de aplicações.

As abordagens mais populares de análise de movimento estão apon-

tando para modelos de probabilidade espaço-temporal (REVATHI; KUMAR, 2012). A incerteza inerente das observações em cenas de vídeo é um problema característico da videovigilância automatizada, o qual reforça o uso de raciocínio probabilístico na modelagem dos eventos. Para tanto, os formalismos de modelagem de máquina mais comuns adotam Redes Bayesianas (BN - Bayesian Networks) (ver subseção 2.2.1), Redes Bayesianas Dinâmicas (DBN - Dynamic Bayesian Networks), Filtros de Kalman (CZYZEWSKI; DALKA, 2008) (KF - Kalman Filters), Modelos Ocultos de Markov (HMM - Hidden Markov Models) incluindo suas variações e Modelos de Misturas Gaussianas (GMM - Gausian Mixture Models). Para o reconhecimento de padrões e treinamento desses modelos, predominam métodos estatísticos de aprendizagem como o de Maximização de Expectativas (EM - Expectation Maximization), Máquinas de núcleo (SVM - Support Vector Machine) (CRISTIANINI; SHAWE-TAYLOR, 2000) e Redes Neurais(NN - Neural Networks) (ZENG; CHEN, 2011).

Alguns exemplos de uso desses formalismos serão explorados ao longo das próximas seções deste capítulo juntamente com as abordagens mais recentes neste campo de pesquisa. A atenção maior vai para os modelos probabilísticos GMM e o método estatístico de aprendizagem EM. Eles são os mais apontados na revisão bibliográfica para o objetivo deste trabalho e portanto foram os alvos de uso para as implementações dessa proposta (FIGUEIREDO; JAIN, 2002; MORRIS; TRIVEDI, 2008; LAVEE et al., 2009; ZHANG et al., 2009).

No escopo de reconhecimento de padrões, algumas formas de aprendizagem podem ser adotadas a partir dos dados observados nas cenas de vídeos como a **aprendizagem indutiva** usando árvores de decisão (RUSSEL; NORVIG, 2009). Essas técnicas usam uma realimentação (ou *feedback*) na apresentação dos dados de entrada (ou textbfexemplos) de modo que para cada passo na busca de uma decisão, alguma heurística associada a seus atributos induza à um caminho mais seguro na solução do problema dentro das inúmeras hipóteses possíveis. Outro tipo de aprendizagem, chamada de **aprendizagem supervisionada**, usa as respostas corretas para cada exemplo como *feedback*. Se esta aprendizagem tem a função de mostrar valores discretos de decisão na sua saída, pode-se chamar isso de **classificação**. A aprendizagem supervisionada portanto, pressupõe um **treinamento supervisionado** do conjunto de dados e é um recurso utilizado na construção de modelos para os sistemas de videovigilância pois resulta em respostas rápidas na classificação de eventos ou de reconhecimento de objetos após uma fase de treinamento prévio. No entanto esse recurso engessa a "inteligência" do sistema e consequente liberdade de aplicações visto que precisam ser alimentados por exemplos fornecidos por algo ou alguém que detém algum nível de

conhecimento do problema.

Por outro lado, há o treinamento **não supervisionado** o qual não requer exemplos para convergir a aprendizagem. Esses são modelos de maior complexidade e contribuem para aquelas aplicações que buscam sistemas de fato autônomos e adaptativos no reconhecimento de padrões anômalos. Entende-se por adaptativo aqueles sistemas que mudam sua "consciência" na medida em que as observações produzem amostras que podem mudar com o tempo. Essas aplicações são possíveis de implementação através de modelamento e técnicas apropriadas quando os dados são aleatórios, crescentes, mutáveis e controláveis.

2.2 MODELAGEM DA ANÁLISE DE VÍDEO

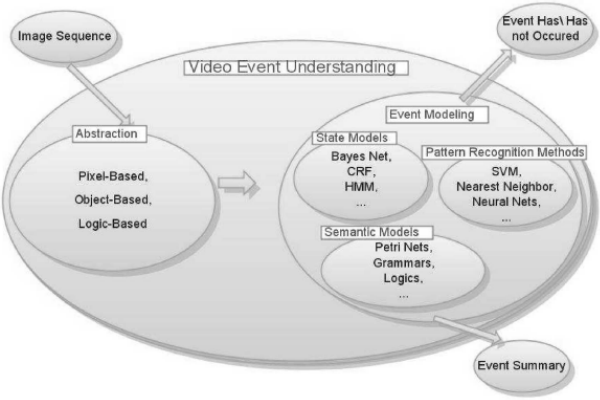
Analisar eventos em cenas de vídeo no espaço e tempo é um processo inerentemente da cognição humana e portanto complexo. Construir este processo em sistemas computacionais requer uma modelagem que se inicia no tratamento do conteúdo de baixo nível dos dados correspondentes às sequencias de imagens e termina em um nível mais alto de abstração, onde pretende-se estabelecer o entendimento sobre a semântica dos movimentos. A modelagem do sistema que irá desempenhar essa tarefa é fundamental uma vez que há caminhos possíveis para alcançar as diferentes metas na análise.

Segundo Lavee et al. (2009) há duas saídas importantes nessas modelagens. Uma indicando se o evento está simplesmente ocorrendo ou não e outra que apresenta uma resposta mais completa sobre o significado do evento observado. A Figura 4 resume esta modelagem destacando que ela precisa ser suportada por algum método de reconhecimento de padrões. O evento pode ser entendido como um aspecto saliente encontrado na abstração de baixo nível das sequências de imagens e que estão ocupando um espaço de tempo relevante na observação. Portanto, cada formalismo dos modelos de estado do evento captura um aspecto importante ou proprietário dos eventos de vídeo.

Movimento, atividade, ação e evento são termos distintos de tratamento na área de videovigilância. Em relação a este trabalho, a meta foi focalizar atenção ao deslocamento no espaço-tempo de objetos móveis no ambiente, criando um modelo que aprende as trajetórias recorrentes. Portanto, foi determinado que a atividade ou ação particular de cada objeto não é relevante para a análise. Desse modo, características do objeto como sua silhueta ou aparência não foram mérito de coleta e tratamento nos dados.

Formalismos de alguns métodos, modelos, ferramentas estatísticas e abstrações ilustrados na Figura 4 estão bem estudados e matematicamente bem formulados e por isso frequentemente aparecem como modelos adotados

Figura 4: Um modelo de análise de vídeo



Fonte: Lavee et al. (2009).

para resolver problemas em rastreamento ou identificação de comportamento. Na revisão bibliográfica sobre o tema, fica evidente que não há um método ou abordagem genérica que consiga abranger uma gama de aplicações.

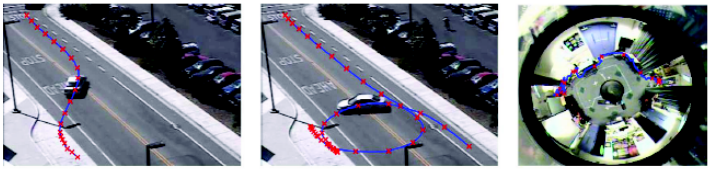
Na pesquisa de Lavee et al. (2009) o processo para extração de características deve passar por uma etapa de abstração baseada na análise de *pixels* (cor, textura, etc), objetos (primitivas de tamanho, forma, trajetória, etc) ou lógica (estados possíveis) das sequências de imagens para então serem utilizadas por uma modelagem do evento. Esse modelo determinará como saída, se um evento de interesse está ou não ocorrendo ou ainda descrevendo mais detalhes sobre este evento, caso a aplicação exija.

A análise de comportamento é um tema bastante amplo dentro da visão computacional. Ela se desdobra em várias linhas de pesquisa e é dependente da aplicação final a qual deseja-se automatizar ou semi-automatizar. O interesse maior nessas linhas de pesquisa são aquelas direcionadas para análise em alto nível. Além da análise do movimento do rastreamento de múltiplos objetos móveis, mérito de estudo da videovigilância e deste trabalho, outras aplicações podem exigir diferentes formas de modelagem com uso particular de um conjunto de ferramentas e métodos específicos. São exemplos disso, como já citado no capítulo anterior a análise do movimento da forma de caminhar (*Gait Recognition*), dos gestos (*Gesture Recognition*), de multidão (*Crowd Analysis*), do corpo humano (*Human Activity*), dos objetos (*Motion Analisys*) entre outros.

Para se exemplificar uma dessas linhas de pesquisa, pode-se ilustrar

o trabalho de Morris e Trivedi (2008) mostrado na Figura 5. São exemplos de trajetórias indicando comportamentos anormais dos movimentos de um automóvel cruzando uma pista na contra-mão (a) ou realizando um loop de 360^o (b) ou ainda de uma pessoa caminhando em um percurso atípico, junto as paredes de uma sala vídeo monitorada com câmera usando lente especial para visão panorâmica (*eye mirror*) (MORRIS; TRIVEDI, 2008). Nesses casos, os objetos alvos são pessoas ou veículos restritos somente na trajetória analisada (Motion Analisys).

Figura 5: Exemplos de comportamentos indicando trajetórias anormais.

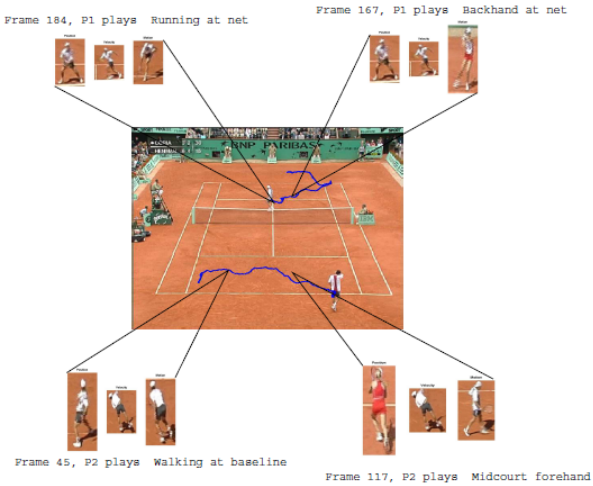


Fonte: Morris e Trivedi (2008).

Um outro exemplo ilustrado na pesquisa de Robertson e Reid (2006) vai além da análise do comportamento da trajetória. Eles exploram um método genérico para reconhecimento de atividade humana (Human Activity) em vídeo descrevendo que o comportamento humano pode ser modelado como uma sequencia estocástica de ações. As ações podem ser descritas como características de um vetor que representa a trajetória (velocidade e posição) e um conjunto de descritores locais do movimento em baixo nível, extraídos de uma janela local do alvo, ou seja do humano em movimento. Usando uma técnica bayesiana para associar as características de ações no espaço e tempo com o respectivo alvo, o autor utiliza modelos ocultos de Markov(ou Hidden Markov Models - HMM) (YU; MOON, 2009) para estimar o comportamento através da associação com um outro modelo baseado em conhecimento sobre jogadas de tênis.

A Figura 6 ilustra a estratégia proposta a qual oferece como saída do sistema, interpretações de alto nível sobre as cenas observadas. Neste exemplo alguns padrões de amostras foram oferecidos no treinamento do HMM. Para cada *frame*, um cálculo de probabilidade com o padrão é realizado para confrontar com os *frames* das cenas observadas. Na tabela da Figura 6 o erro da interpretação foi destacado com o texto escrito em itálico. As jogadas estimadas pelo sistema que não estão grifados em itálico representam sucesso na sua interpretação.

Figura 6: Exemplo de interpretação de atividade humana no ambiente.



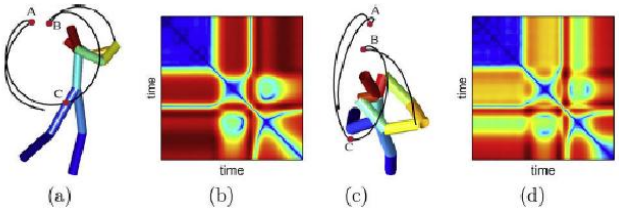
Frame	Estimated stroke	Likelihood	Ground truth
32	P2: *Baseline forehand*	0.41	Serve
39	P1: Walking at baseline	0.54	Walking at baseline
51	P1: *Running at baseline*	0.60	Running at mid-court
76	P2: Walking at baseline	0.36	Walking at baseline
99	P2: Walking at mid-court	0.45	Walking at mid-court
132	P2: Backcourt forehand	0.66	Backcourt forehand
148	P1: *Forehand at net*	0.34	attempted forehand

Fonte: Robertson e Reid (2006).

Ainda nessa linha de trabalhos encontramos o artigo de Poppe (2010) que usa as chamadas **matrizes de auto-similaridades**. Elas são usadas para identificar similaridades de ações na análise de comportamento humano, mas agora sobre seu próprio movimento. As matrizes são uma representação gráfica de sequências similares em uma série de dados, os quais resultam em diferentes medidas como distancia espacial, correlação e a comparação de histogramas locais dos objetos em movimento. Esta técnica visa buscar um determinado padrão em uma série de dados que é o registro de um determinado comportamento. Portanto comportamentos similares produzem matrizes similares fracamente dependentes do ponto de vista da câmera. A Figura 7 ilustra um exemplo desta técnica para a simulação de uma pessoa realizando uma tacada de golf. A movimentação em pontos de vista diferentes

(a e c) geram padrões muito semelhantes da matriz de similaridade calculada sobre os pontos de referência A, B e C (b e d) (POPPE, 2010).

Figura 7: Exemplo de interpretação de atividade humana sobre seu próprio movimento.

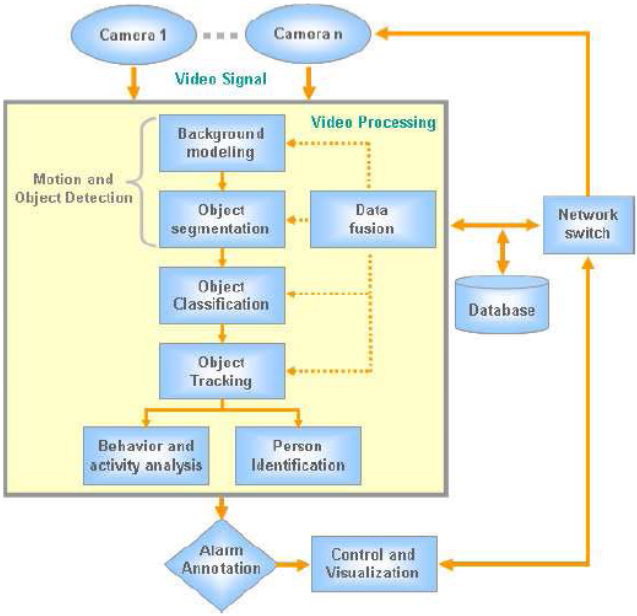


Fonte: Poppe (2010).

Nesta ideia, é possível construir padrões de comportamento no rastreamento de um humano em um determinado ambiente que possui pontos de referência bem definidos. O traço de deslocamento similar implica em comportamentos similares. Um dos desafios neste caso é o tratamento da grande dimensão de dados da imagem extraída e consequente custo computacional nas etapas do processamento dos *frames* de vídeo que compreendem a captura, levantamento e análise dos descritores da silhueta dos humanos (objetos) observados. Uma técnica comum de redução destas dimensões de dados, é a PCA (Principal Component Analisys) que deve ser utilizada com critérios caso existam ruídos na aplicação (ZHANG et al., 2009) (YU; MOON, 2009). Outras opções foram melhor relatadas por Poppe (2010) como LLE (Local Linear Embedding), LPP (Localy Preserving Projections) e LSTDE (Local Spatio Temporal Discriminant Embedding).

Examinando o trabalho de Ko (2008), que acontece depois de quase 5 anos após a avaliação da revisão de trabalhos relacionados na área, feita por Hu et al. (2004), parece existir um consenso sobre como um sistema de videovigilância automatizado deve ser modelado, especificamente se ele é baseado no rastreamento dos objetos no cenário. O *framework* genérico mostrado na Figura 8 ilustra os principais processos envolvidos nesses sistemas. As mudanças nos últimos anos estão associadas aos recursos tecnológicos mais acessíveis que permitem expansões desses sistemas em uma arquitetura de rede incluindo a criação de bases de dados em todos os níveis de processamento das imagens, como aqueles comentados sobre a quarta geração dos sistemas de videovigilância, na seção 1.1.2. Independente desses avanços, ainda há um conjunto de processos envolvidos no processamento de vídeo que continua o mesmo.

Figura 8: Um *framework* genérico de um sistema de videovigilância automatizado.



Fonte: Ko (2008).

Detalhando um pouco mais sobre o *framework* da Figura 8, todo o processamento e análise de vídeo recebe *frames* oriundos de uma ou mais câmeras e a partir daí, métodos são utilizados para detectar (segmentar) movimento e os respectivos objetos presentes na cena. O sucesso das funções dos próximos estágios é extremamente dependente dessa etapa. Portanto, a modelagem precisa incluir as melhores estratégias, métodos ou técnicas possíveis para reconhecer e extrair com fidelidade o fundo estático da cena permitindo assim a segmentação dos objetos e respectivos movimentos. A propósito, esse ainda é um tema que tem se apresentado como um desafio, gerando trabalhos onde a diferenciação temporal, subtração do fundo de cena (detecção de *foreground*) e fluxo ótico são técnicas mais frequentes (KO, 2008). Essas abordagens são direcionadas para localizar regiões de *pixels* que representam objetos móveis dentro da cena.

O método de diferenciação temporal é um método simples que realiza a diferença entre dois ou três *frames* consecutivos para encontrar as regiões

móveis. Essa ideia tem limitações de uso pois não capta todos os *pixels* relevantes. Já o método de subtração do fundo de cena é mais popular especialmente para aquelas aplicações que possuem fundos de cena pouco complexos, ou seja, mais estáticos e com limitada diversidade de texturas, objetos e contrastes. Este fundo de cena é tomado como referência e é subtraído do *frame* atual. Adotando-se um limiar de referência para essa subtração, o resultado será uma matriz binária que destaca a localização dos *pixels* relacionados com os objetos móveis de primeiro plano (*foreground*) para cada *frame*. No entanto, em quase todos os ambientes reais monitorados por câmeras, o fundo de cena vai sofrendo modificações ao longo do tempo e como consequência, um outro desafio para resolver neste tipo de abordagem. Técnicas estatísticas tem sido utilizadas para contornar esses problemas, decidindo em função de probabilidades, se cada *pixel* da imagem faz parte ou não do *foreground*.

O método de fluxo ótico descreve os deslocamentos ocorridos entre dois *frames* consecutivos. O campo de velocidade gerado é frequentemente descrito no domínio discreto através de um mapeamento vetorial conhecido como vetores de deslocamento. Esses vetores determinam com boa precisão a segmentação do objeto e do movimento (GONZALEZ; WOODS, 2008). No entanto são métodos sensíveis a ruídos, oclusão e computacionalmente mais complexos.

Com destaque nessa tarefa de extração do *background*, o trabalho de Barnich e Van Droogenbroeck (2011) desponta em citações de artigos relacionados em função do superior desempenho alcançado pelo ViBe - Visual Background Extractor. Trata-se de um algoritmo que atua sobre cada *pixel* e sua vizinha usando atualizações com mecanismos heurísticos e métodos não paramétricos como cálculo de distâncias euclidianas. Os valores dos *pixels* são constantemente atualizados somente nas regiões onde há movimento a ser segmentado e a inicialização do algoritmo, ao contrário da grande maioria das propostas anteriores, utiliza somente o primeiro *frame* como base.

Continuando na análise do fluxo de processos da Figura 8, a etapa de classificação de objetos cumpre um papel importante para garantir a robustez tanto do rastreamento quanto da análise do comportamento (ELHOSEINY et al., 2013). Em geral, a classificação é realizada por alguma metodologia de aprendizagem como por exemplo Redes Neurais NN - (Neural Networks) e SVM - Support Vector Machine. Elas se baseiam na forma dos objetos a partir de um vetor de características (descritores) como pontos, silhuetas ou conjuntos de *pixels* conectados (**blob**). A classificação também pode se basear em métodos mais elaborados como por exemplo a partir da representação do movimento característico e periódico de uma pessoa ou a partir do histograma local do objeto.

Para a etapa de rastreamento de múltiplos objetos, a classificação prévia

pode ser dispensável uma vez que há dois fundamentos básicos para o rastreamento: A segmentação dos objetos a partir do *foreground* e a correta identificação de cada um deles ao longo do fluxo de *frames*. Portanto, para identificar a correspondência entre n objetos e os correspondentes n traços de deslocamento, existem $n!$ atribuições possíveis. A tarefa torna-se mais complexa se oclusões parciais ou totais ocorrem ou ainda se os objetos saem do FOV ou migram para FOVs de outras câmeras. Em geral, utilizam-se os dados gerados pelo próprio rastreamento ou por um ou mais processos nesse encadeamento visando tornar a associação entre os objetos e os respectivos rastros mais confiável (robusto). Esse problema também chamado de **associação de dados** (MA; WAN, 2009), merece ferramentas estatísticas para ser controlável, evitando torná-lo uma tarefa NP-Difícil (RUSSEL; NORVIG, 2009) por conta da fusão de dados originada nas várias etapas do framework, ou mesmo externos à ele.

Na última etapa do *framework* de Ko (2008) reside um dos desafios mais estudados no domínio da visão computacional e da inteligência artificial que é o entendimento e aprendizagem de comportamento semântico a partir de atividades observadas em uma vídeo monitoração. Segundo o autor e Hu et al. (2004), confirmadas pela revisão das demais referências relacionadas avaliadas neste trabalho, muitas propostas tem sido apresentadas sobre os níveis mais baixos do processamento de imagem desde a detecção de objetos até o rastreamento, no entanto poucas tem explorado com confiabilidade a classificação e entendimento de atividades dos objetos, especialmente os humanos.

De fato, detectar comportamentos suspeitos ou intenções hostis de pessoas exige um modelamento apropriado e carregado de regras devido a aleatoriedade e complexidade da natureza do movimento ou intenções humanas. A abordagem realizada neste trabalho se distancia do propósito de identificar movimentos neste nível de abstração pois nem a semântica e nem o contexto estão sendo levados em consideração no modelamento. Mesmo preocupando-se somente com as anomalias de deslocamentos locais (próxima localização) ou globais (próximas n localizações) ainda é possível inferir sobre anormalidades de comportamento pois más intenções ou atitudes suspeitas podem estar relacionadas com "movimentação não usual" dos objetos monitorados. Obviamente que realizando uma fusão de dados a partir de sensores adicionais no ambiente ou ainda a partir do reconhecimento de padrões como gestos, expressões faciais ou outras disposições emocionais, teríamos uma inferência sobre o movimento mais precisa, porém isso não é o foco do presente trabalho.

Neste sentido, apesar dos autores apresentarem várias outras técnicas para modelar a última etapa do seu *framework* geral e que também se en-

contram exemplificadas no modelo de Lavee et al. (2009) da Figura 4, como FSM - Finite State Machine e Grammatical Techniques, as abordagens mais populares ainda são de modelos estatísticos. Aspectos conceituais gerais da maioria dessas técnicas não-probabilísticas, podem ser consultadas em Russel e Norvig (2009).

2.2.1 Modelos de Probabilidade Espaço-Temporal

A representação probabilística dos relacionamentos entre variáveis é um atributo da teoria da probabilidade que permite captar incertezas do conhecimento de modo natural. Então, sistematicamente podemos representar esses relacionamentos elegendo as variáveis que melhor representam o estado do mundo, ou seja, da aplicação. Pode-se então por exemplo construir grafos acíclicos orientados que representam a evolução dos estados (nós pai) para outros estados (nós filho) por vínculos (probabilidades condicionais) que melhor representam um conhecimento nessas transições de estados. Este é o caso da formação de redes Bayesianas que podem ajudar a inferir resultados a partir de suas entradas, mas somente para um cenário onde as variáveis não mudam com o tempo. Quando o tempo é outra variável a ser considerada novas hipóteses precisam ser consideradas de modo que o estado atual seja representado somente pelos vínculos de estados imediatamente anteriores para reduzir o número de pais do nó atual e consequente complexidade da análise. Então, se só tomarmos o estado anterior como base, estaremos realizando uma análise de primeira ordem e se considerarmos outros estados antecessores consequentemente a ordem da análise aumenta. Como o foco deste trabalho trata da análise de movimento, o tempo é uma variável fundamental para a caracterização do seu entendimento. Nesses casos, os modelos de probabilidade espaço-temporal conseguem descrever a evolução dos estados ao mesmo tempo que descrevem o processo de sua observação. Portanto, são os modelos mais usuais como ferramentas para serem usados não só para análise de movimento como para o rastreamento de objetos móveis.

2.2.1.1 Os Modelos Ocultos de Markov - HMM

Os modelos HMM e suas variações estão entre os mais populares (LAVEE et al., 2009; YU, 2010) nas modelagens de eventos de videovigilância, especialmente pelas suas propriedades de combinar a modelagem da evolução temporal com a modelagem probabilística dos eventos ou estados. O HMM clássico tem uma estrutura gráfica particular que descreve um

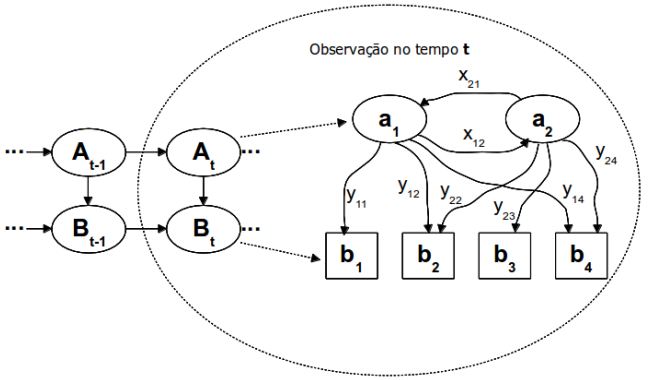
modelo onde as observações atuais são dependentes apenas sobre o estado atual. A situação atual é apenas dependente sobre o estado no intervalo de tempo anterior. Uma vez que a estrutura HMM é fixa e repetitiva, podemos definir a probabilidade de longa sequência de estados, especificando um conjunto de parâmetros em um número tal que depende dos estados possíveis e os símbolos de observação.

O modelo oculto de Markov é estatístico e paramétrico. Um sistema modelado é assumido como um processo de Markov com parâmetros desconhecidos, e o desafio é determinar esses parâmetros desconhecidos (ocultos) a partir dos parâmetros observáveis. Extraindo-se os parâmetros deste processo consegue-se então utilizá-los para realizar novas análises em um processo contínuo e temporal. Este modelo portanto é extremamente útil para aplicações de reconhecimento de padrões, como a fala, escrita, gestos, atividade humana e inclusive o reconhecimento de padrões de movimento de objetos.

Em um modelo regular de Markov, o estado é diretamente visível ao observador, e portanto os únicos parâmetros usados são as probabilidades de transição de estado. Cada estado possui uma distribuição de probabilidade sobre os possíveis resultados. A Figura 9 ilustra um exemplo de modelo oculto de Markov de primeira ordem. Cada elemento no círculo representa uma variável aleatória que pode também possuir várias outras formando uma megavariável (YU, 2010). A variável aleatória $A_t \in \{a_1, a_2, a_3 \dots\}$ é o estado oculto no instante de tempo t. A variável aleatória $B_t \in \{b_1, b_2, b_3 \dots\}$ representa as variáveis observáveis no instante de tempo t. As setas distribuídas no diagrama indicam as dependências condicionais entre as variáveis ocultas e observáveis. A partir do diagrama pode-se concluir que o valor da variável oculta A_t depende exclusivamente do valor da variável oculta A_{t-1} no instante de tempo $t-1$ (dependência de primeira ordem). Esta é a hipótese de Markov que afirma que o estado atual depende apenas de um histórico finito de estados anteriores. Nesta mesma hipótese, o valor da variável observada B_t depende exclusivamente do valor da variável oculta A_t, ambas no instante de tempo t. Na Figura 9 os valores x_{ij} da evolução temporal de um modelo de Markov representam as probabilidades de transições de estado entre os estados ocultos a_i. Da mesma forma os valores y_{ij} representam as probabilidades de saídas para os estados observáveis b_i.

Um uso comum de HMMs em eventos de vídeo é baseado em um modelo definido por símbolos de observação relacionados com o esquema escolhido para abstração. Os estados do HMM são em geral abstratos, e seu número é escolhido empiricamente. Os parâmetros do modelo HMM podem ser aprendidos a partir de dados de treinamento ou especificado manualmente usando o domínio de conhecimento do evento. Para identificar eventos dife-

Figura 9: Exemplo de um modelo HMM.



Fonte: Elaborado pelo autor.

rentes, o mesmo procedimento modelo é treinado para cada caso em especial. Exemplos teste são então avaliados para determinar qual a probabilidade de cada evento ter sido gerado a partir de cada um dos modelos HMM. O evento modelo que produzir a maior taxa de probabilidade, é então usado para rotular o teste exemplo.

O modelo HMM tem sido estendido em várias maneiras para adaptar os desafios da modelagem de eventos de vídeo. São topologias dos nós da rede que são rearranjadas em maneiras diferentes para se conseguir resultados mais significativos e dependentes da sua aplicação. Alguns exemplos foram descritos por Lavee et al. (2009). Outros trabalhos como o de Xiang e Gong (2005) adotam modelos estendidos de HMM como o chamado de MOHMM - Multi-Observation Hidden Markov Model. Nesse caso os autores tem o objetivo de identificar perfis de comportamento e anomalias sem que seja necessário rotular a sua base de dados de vídeo de treinamento, ação comum em muitas abordagens como realizado em (BERCLAZ et al., 2008). Para tanto o número de estados ocultos para cada variável do modelo HMM é associado a quantidade de perfis aprendidos pelos agrupamentos de amostras observadas. O modelo vai aprender um MOHMM para cada perfil e então, objetos com movimentos fora desses perfis, vão gerar um valor de probabilidade pequeno. Em função disso esses movimentos irão ser detectados e rotulados como anormais.

Outra variação de HMM é abordada por Yu (2010), o qual descreve com detalhes o uso de HSMM - Hidden Semi-Markov Models, destacando

o gradual uso dessa extensão de HMM como um modelo apropriado nas aplicações de reconhecimento de atividade humana na vida diária em pequenos ambientes. Como uma atividade principal, atividades usuais como gastar tempo em frente a uma geladeira ou a movimentação de humanos entre locais designados, foram modelados como uma variável para cada estado do HSMM que representando assim uma atividade **atômica** onde sua duração representa o tempo de ação atômica modelado.

A modelagem com HMM e suas variações são apropriadas para diversas aplicações, mas exigem modelos para cada comportamento usual que precisam ser modelados adequadamente e devidamente treinados. A dinâmica dos cenários observados na vida real coloca grandes desafios para o uso desse tipo de modelagem, especialmente se desejamos aplicá-las em problemas que não possuam qualquer treinamento prévio.

2.2.1.2 Modelos de Misturas Gaussianas - GMM

Em reconhecimento de padrões estatísticos inerente da análise de movimento, agrupamentos finitos de misturas de dados amostrados permitem uma abordagem de aprendizado não supervisionado, como definido na seção 2.1. Essas misturas representam observações que foram produzidas durante o rastreamento *frame* a *frame* de um objeto móvel e que portanto são aleatórias.

Os objetos móveis, quando em uma abordagem baseada em rastreamento explícito, produzem uma quantidade significativa e constante de dados não supervisionados que categorizam sua trajetória através de seus atributos discretos ou contínuos. No caso de existirem vários atributos que representam os objetos móveis ou traços de deslocamento, a escolha natural da função que melhor representa as distribuições de probabilidade entre os atributos é a distribuição gaussiana multivariada. Basta um atributo desses ser contínuo. No caso de aplicações em videovigilância o atributo contínuo *tempo* já garante o uso desse tipo de distribuição. Essas distribuições são chamadas de misturas de distribuições gaussianas e em geral são finitas e multivariadas. Métodos como o GMM são capazes de representar complexas funções de densidade de probabilidade (*pdf - probability density function*) como essas.

Os componentes de um vetor são variáveis que devem representar bem o modelo de qualquer etapa em um *framework* como o apresentado na Figura 8. O número de variáveis vai determinar a dimensão dos planos de distribuição de *pdf* e naturalmente o custo computacional para isso. Se o vetor de características estiver bem modelado, as amostras capturadas durante a fase de treinamento *frame* a *frame* vão gerando de forma não supervisionada, formação dispersa de pontos inter-relacionados que podem ser discriminados

através de agrupamentos com o auxílio do algoritmo de aprendizagem EM que será visto na próxima seção. Os autores Panda e Meher (2013) desenvolveram uma proposta de aplicabilidade de um GMM para auxiliar na tarefa de detecção de *foreground* em sistemas de câmeras estacionárias. Eles usaram uma taxa de aprendizagem de um peso Gaussiano para ajustar os parâmetros do modelo com base na extração de fundo de cena através da exploração da correlação de vizinhança de um pixel.

Tanto o HMM quanto o GMM, são modelos definidos como parametrizados. Eles precisam de um treinamento do conjunto de amostras espalhadas normalmente em hiperplanos do espaço n-dimensional por conta da n-dimensionalidade dos vetores que caracterizam as variáveis de interesse (objetos ou traços de deslocamento em nosso caso). Para tanto o algoritmo EM é o mais comum nessa tarefa. Dependendo da aplicação que se quer modelar nem sempre existe um modelo ou método perfeito para se adotar. As desvantagens que podem surgir acabam exigindo algumas condições de contorno para levar estes modelos a responderem com os melhores resultados possíveis. O HMM por exemplo necessita a criação de modelos de estados e transições antes de realizar seu treinamento e ainda precisa ter esse processo repetido se novas amostras necessitarem entrar no treinamento. O GMM não tem esse problema mas dados esparsos podem dificultar a discriminação de agrupamentos e por consequência sua convergência. Variações, adequações e combinações dessas ferramentas com outros métodos heurísticos ou estatísticos são exploradas pelos autores multiplicando abordagens para as mesmas aplicações.

Propostas como a de Xiang e Gong (2005) usam o GMM tanto para segmentar os objetos quanto para representar comportamentos baseados em evento. Nessa modelagem o autor utilizou um vetor de 7 dimensões: a posição 2D do *blob*, a dimensão *blob (w,h)* e mais três variáveis associadas a modelagem da segmentação de movimento onde foi usado a técnica de PCH - Pixel Change History. Na mesma linha a proposta de (BASHARAT et al., 2008) usa um vetor de 5 dimensões para caracterizar cada objeto móvel da cena com a posição do **centroide**[1] do seu *blob* no *frame*, a dimensão do *blob (w,h)* e o valor de tempo *t* associado ao seu deslocamento.

2.2.2 Modelos de Aprendizagem e Treinamento

Uma outra classe de aprendizagem para tratar volumes de dados que dados são aleatórios e crescentes é chamada **aprendizagem estatística**. Esse

[1] Centroide é definido aqui como o centro geométrico da forma retangular com altura e largura que circunscreve um objeto sob análise.

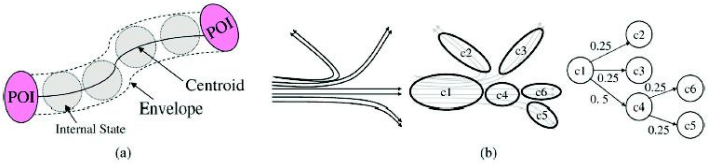
tipo de aprendizagem pode fazer uso do treinamento de parâmetros de funções de distribuição de probabilidade *pdf* que melhor representam o comportamento estatístico desses dados. Por exemplo estes parâmetros podem ser a média e variância dos dados de funções gaussianas, comuns para essas aplicações. Nesse caso trata-se de uma *aprendizagem paramétrica* que é útil para qualquer espaço n-dimensional. A aprendizagem estatística também pode pode usar heurísticas para determinar a separação ideal de agrupamentos que representam padrões. Esta aprendizagem é chamada de *aprendizagem não-paramétrica* e inclui um extenso grupo de modelos. Os mais presentes dentro do propósito deste trabalho são as redes neurais (ZENG; CHEN, 2011; TEHRANI et al., 2009), os modelos de vizinhança como o *K-nearest*, os modelos de núcleo como o *K-means* (HU et al., 2006) e as máquinas de núcleo como o SVM (SUDO et al., 2008). Ainda entre esses modelos, o SVM tem um visível destaque dentro de linhas de pesquisa que buscam robustez especialmente na etapa de rastreamento de objetos nas cenas de vídeo (PONTIL; VERRI, 1998; CRISTIANINI; SHAWE-TAYLOR, 2000).

Uma vez que a estrutura do modelo é especificado para uma dada aplicação, os parâmetros devem ser aprendidos a partir de dados apresentados na fase de treinamento e assim o modelo está pronto para a fase de teste de desempenho (ou monitoração). Este é um procedimento usual de testar os modelos sobre base de dados de referência que estão disponíveis através de sequências de vídeo para treinamento, e em outras sequências, geralmente em menor número para teste. As sequências de vídeo para o treinamento estão com conteúdos que não apresentam anomalias de movimentos dos objetos em cena para permitir que os modelos criados produzam os dados relativos a comportamentos normais das cenas. As sequências de testes possuem de forma aleatória a presença de movimentos não usuais que devem ser detectados pelos modelos. O desempenho do modelo então é medido em função da relação entre a quantidade de falsos positivos encontrados contra o número total de anomalias reais contidas nas cenas.

Um ponto importante dentro dos esquemas de aprendizagem é como modelar este processo. Segundo Morris e Trivedi (2008) o modelamento de trajetórias tem sido realizado de duas maneiras conforme os esquemas ilustrados na Figura 10. Em (a) o Ponto de Interesse - POI (a) circula entre áreas da cena carregando informação que deve ser amostrada ao longo do tempo usando na forma mais simples a posição de seu centroide, ou a média de posições com sua variância formando um envelope do caminho atravessado por regiões que podem ser modeladas como estados internos de observação da presença do objeto. Neste esquema tanto Modelos GMM ou HMM fornecem eficientes formas de aprendizagem dos caminhos (MAGGIO; CAVALLARO, 2009). Já a Figura 10 (b) mostra que os caminhos observados foram

representados como uma árvore de estados já computando os seus valores de probabilidade. Este esquema também é ideal para ser usado com redes Bayesianas ou HMMs. A criação dos estados bem definidos das observações ($C1 \cdots C5$) pode ser feita manualmente ou automaticamente, rotulando-se estas regiões na área da cena com estratégia de aprendizados de agrupamento de misturas gaussianas. Propostas que trabalham com contexto fazem uso dessa estratégia (MAGGIO; CAVALLARO, 2009; LI et al., 2012b).

Figura 10: Exemplos de modelagem de trajetos.



Fonte: Morris e Trivedi (2008).

A detecção de anormalidade de um movimento, que implica em um comportamento anormal de um objeto móvel da cena, é consumado desde que o caminho tomado por tal objeto não se acomode bem aos caminhos típicos já aprendidos (MORRIS; TRIVEDI, 2008). Em uma expressão pode-se dizer que se a maior probabilidade do caminho de um objeto i, por exemplo λ^i dado uma trajetória F for menor que um limiar L_{λ^i} implica inferir que objeto está seguindo um padrão anormal. Ou seja, se $p(\lambda^i|F) < L_{\lambda^i}$ o movimento é considerado anormal. O valor do limiar L_{λ^i} pode ser ajustado para cada caminho e individualizado para cada objeto segundo o que se aprendeu no modelo adotado no treinamento dos caminhos ativos. Esta tem sido a ideia geral utilizada como saída dos sistemas que vão responder sobre comportamentos atípicos dos objetos.

2.2.2.1 Aprendizagem Estatística com Algoritmo EM

A escolha do treinamento em função da modelagem escolhida é fundamental para a confiabilidade do sistema. Dentre os vários tipos de algoritmos de treinamento não supervisionado, o Expectation-Maximization (EM) é o mais citado para treinar os modelos estatísticos na análise de movimento (FIGUEIREDO; JAIN, 2002; XIANG; GONG, 2005; SUDO et al., 2008).

As amostras geradas nas misturas gaussianas não possuem rotulação de categoria pois não são avaliadas por qualquer supervisionamento. Cada

amostra está associada a categorias diferentes das variáveis que formaram o número de agrupamentos, chamados de **componentes**. Cada componente da mistura possui os parâmetros gaussianos μ_i (média), Σ (co-variância) e w (peso = sua probabilidade na mistura) . A função do método padrão do EM é ajustar interativamente as finitas misturas gaussianas para convergir na probabilidade máxima estimada dos parâmetros misturados de cada componente. Existindo a convergência na fase de treinamento dos dados, cada amostra coletada na fase de testes vai ser associada a um dos componentes da mistura. Esses componentes serão a representação das variáveis determinadas na fase de modelagem do sistema.

Na versão padrão do EM não se conhece os componentes e nem os seus parâmetros. Para inicialização é necessário então fornecer o número de componentes e arbitrar os seus valores de μ, Σ, e w. Logo após calcula-se qual a probabilidade de cada ponto de dados (amostra) pertencer a cada componente. Na posse das melhores probabilidades calculadas, reajusta-se todos os dados aos seus componentes onde cada componente é ajustado ao conjunto completo dos dados. Cada dado então é ponderado pela probabilidade de pertencer a cada componente. Na realidade está se deduzindo distribuições de probabilidades sobre cada componente que representa as variáveis ocultas do sistema modelado.

Após este processo inicial, executa-se a etapa **E - Expectation** que usa a regra de Bayes para calcular a probabilidade de que cada dado tenha sido gerado por cada componente. Ou seja $p_{ij} = P(C = i|x_j)$ (RUSSEL; NORVIG, 2009). Esta etapa representa a expectativa de que o dado x_j foi gerado pelo componente C_i. Para isso relaciona-se as *variáveis ocultas indicadoras* com um valor *1* se isso é verdade ou *0* caso contrário. Por último executa-se a etapa **M - Maximization** onde, de posse dos valores de p_{ij}, $p_i = \Sigma_j p_{ij}$, x_j recalculam-se os valores de μ_i, Σ_i e w_i. Ou seja, esta etapa encontra os novos valores que maximizam a probabilidade dos dados, em função dos valores esperados das variáveis ocultas representadas pelos seu componentes. As etapas E e M devem se repetir até a convergência dos parâmetros de todos os componentes. A quantidade de iterações vai depender de uma série de fatores como a inicialização e a distribuição de dados.

Nesse método padrão encontram-se sérias desvantagens. Ele exige uma inicialização, preferivelmente supervisionada. Também, como ele é um método *guloso* e local e a função de misturas gaussianas é multimodal (que possui vários mínimos locais), a inicialização pode levar convergências para mínimos locais ou pior ainda não convergir caso um componente se restringir a um único ponto de dados. Neste caso a variância é zero e a probabilidade tende a infinito. Outro problema é a necessidade de se conhecer o número de componentes, que acaba engessando para uso em aplicações que reque-

rem treinamento totalmente não supervisionado. Por conta da recursividade o custo computacional cresce com o número de amostras e componentes.

Mesmo assim, na busca de opções de algoritmos para compor a abordagem proposta aqui, encontra-se alguns autores que apresentam propostas para tornar o EM insensível as desvantagens apontadas até porque, além do algoritmo ser adequado para aplicações que envolvem a formação de agrupamentos com a utilização de misturas de gaussianos, ele também se acomoda bem para aprendizagem de redes bayesianas e HMM.

É o caso do artigo de Figueiredo e Jain (2002). Trata-se de um ótimo referencial para entender melhor os problemas do EM e como contornar suas desvantagens quando ele é utilizado na sua forma padrão. Os autores apresentam um algoritmo que é capaz de selecionar o número de componentes (de agrupamentos) de forma automática (não supervisionada) e sem a necessidade de cuidados na inicialização de dos modelos de misturas finitas a partir de dados multivariados. O contorno destas desvantagens permite ampliar o uso deste método de aprendizagem naquelas aplicações que exigem independência de treinamento ou rotulação prévia dos dados para permitir verdadeiramente seu uso em modo on- line, adaptativo e não supervisionado, caso buscado nesta proposta.

2.2.2.2 Aprendizagem Estatística com SVM

Máquinas de Vetor Suporte, ou Support Vector Machines (SVMs), são também úteis para o aprendizado computacional. São baseadas na teoria de aprendizado estatístico tendo como ideia principal o mapeamento do espaço de estados dos dados de entrada para um outro espaço onde se determine um hiperplano que os separe linearmente. SVMs mostram um desempenho relevante nas aplicações de reconhecimento de imagens e reconhecimento de padrões de comportamento (ELHOSEINY et al., 2013). Diferente de redes neurais, SVMs não processam busca por múltiplos mínimos locais mais sim por um máximo global. Por ser um método classificador muito bem suportado por matemática e estatística ele possui uma boa capacidade de generalização uma vez que ele consegue classificar dados que não pertençam, ao conjunto utilizado em seu treinamento. Isso também leva este tipo de método ser robusto para objetos com grandes dimensões de dados (CRISTIANINI; SHAWE-TAYLOR, 2000).

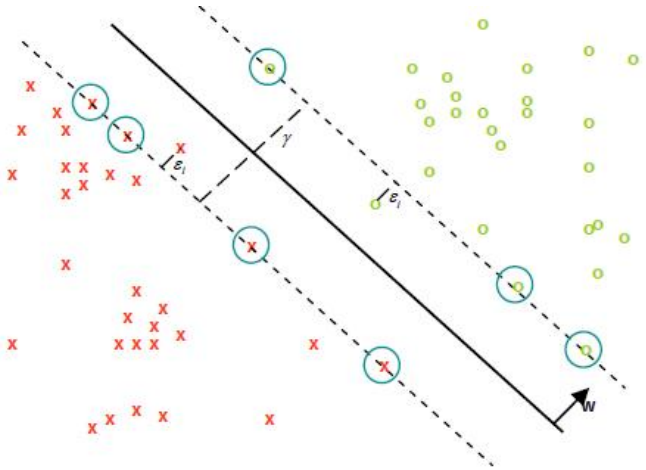
Em síntese, suponhamos inicialmente que nossos dados de treinamento sejam linearmente separáveis, ou seja, que há um hiperplano y na direção w tal que todos os dados de treinamento classificados como positivos $y_i = +1$ fiquem de um lado do hiperplano e os dados classificados como negativos

$y_i = -1$ fiquem do outro lado do mesmo hiperplano. Máquinas de vetor de suporte procuram maximizar a margem do hiperplano separador usando como base a expressão (2.1):

$$MAX_{w,\gamma}\left\{\gamma - C\sum_{i=1}^{l}\varepsilon_i\right\} \tag{2.1}$$

onde C é um fator de penalização para todo x_i dentro da faixa γ na direção w. Tais pontos definem o hiperplano separador e sua remoção ou deslocamento pode afetar a solução do problema pois pode mudar o valor da margem. A Figura 11 ilustra o problema para duas classes linearmente separáveis (duas dimensões), com os vetores suporte associados (pontos circulados acima e abaixo de w):

Figura 11: Plano separador dos vetores de suporte.



Fonte: Gong et al. (2011).

Tanto redes neurais quanto as SVM têm o mesmo objetivo que é o de achar relações entre os dados que minimizam o erro da classificação (LI et al., 2009). Redes neurais tentam minimizar o risco empírico, isto é, reduzir a probabilidade de erro na classificação dentro dos dados de treinamento. Por sua vez, as SVMs foram criadas para minimizar o erro estrutural, de tal forma que a probabilidade de erro de classificação dos dados, que não são de treinamento, seja minimizada, com isso, elas aproximam melhor a função de classificação ideal. Enquanto em redes neurais, o treinamento é feito de

uma forma iterativa, onde cada passo tenta obter continuamente melhores resultados no ajuste da função de classificação, na maioria dos casos, torna-se difícil determinar quando finalizar o processo iterativo, ou especificar quando se chega a uma boa aproximação (LI et al., 2009; ZENG; CHEN, 2011).

As SVMs tratam os dados de forma simultânea e global, proporcionando uma solução ótima quando o treinamento finaliza. Ainda comparativamente, nos treinamentos de redes neurais há uma auto-modificação, alterando os pesos nas arestas conectadas. Durante esse processo de atualização, a solução ótima pode não chegar ao ponto ótimo global, pois o método pode parar num ponto crítico local. O principal motivo é que os dados de treinamento são alimentados na rede um após o outro e um subconjunto dos dados, podendo gerar uma grande influência nos pesos das arestas, desprezando a contribuição dos demais subconjuntos. Por outro lado, as máquinas de vetor suporte possuem, necessariamente, um ótimo global para um processo de classificação.

No reconhecimento de atividade humana, os dados do plano devem representar os diversos pontos de interesse analisados nas sequencias de imagens. Vários hiperplanos vão separar geograficamente determinados traços de deslocamento dos humanos no ambiente que serão representados por um vetor x_i neste espaço. Comportamentos com elevado grau de similaridade ficarão separados pelos hiperplanos identificando comportamentos normais. Um ponto x_i demasiadamente fora deste espaço implica em uma situação anormal e pode então ser alertada. Uma melhor fundamentação e formulação do SVM bem como outros métodos de aprendizado baseados em núcleo podem ser encontrados em Cristianini e Shawe-Taylor (2000).

2.3 ABORDAGENS NA DETECÇÃO DE MOVIMENTO ANORMAL

Como ilustrado na Figura 3 da seção 1.1, há dois ramos de abordagens para DMA. Aqueles baseados nos dados gerados pelo movimento de todos os objetos sobre o fundo estático da cena e aqueles que se baseiam nos dados gerados pelo rastreamento de cada objeto móvel. Para essas duas abordagens, a quantidade de combinações entre técnicas, modelos, estratégias e ferramentas disponíveis se multiplica especialmente por conta da contribuição dada pelos resultados na pesquisa de processamento de imagens, desde o tratamento até a sua segmentação. Essa evolução fica evidente na área de videovigilância pois tais abordagens tem procurado atender diversos tipos de cenários com fundo de cena complexo *indoor* ou *outdoor*, maior número de objetos e tipos, maior número de câmeras fixas ou móveis, menor consumo de memória ou

tempo de processamento, comportamentos complexos, entre outros desafios.

Na pesquisa de vários trabalhos dentro dessas abordagens, se destacam duas formas de análise dos dados coletados: as baseadas nos dados relacionados a regiões pré-definidas dentro da ROI (TZIAKOS et al., 2010) e as baseadas nos dados relacionados as mudanças da informação dos *pixels* no *frame* (LIU et al., 2010; FEIZI et al., 2012).

As abordagens baseadas tanto em rastreamento quanto em movimento possuem um espectro de trabalhos que propõe detectar movimento anormal levando em consideração a análise de uma porção menor de dados localizados em sub-regiões resultantes de uma **grade fixa** normalmente formada por uma divisão uniforme e fixa da ROI. Outro tipo de grade é aquela onde as sub regiões podem variar em forma e tamanho de acordo com a distribuição de dados na ROI. Essa última é definida aqui como **grade adaptativa**. Como forma exclusiva de análise do presente trabalho, foi realizado a análise sobre um novo tipo de grade a qual se ajusta a densidade de dados dentro do ROI, aqui definida como **grade móvel**.

Esta seção destaca alguns trabalhos que representam bem essas duas vertentes de estudo, em especial aqueles onde a análise de movimento anormal é baseada no rastreamento explícito dos objetos móveis, mérito desta tese. Embora sejam abordagens distintas, elas revelam estratégias comuns quando o assunto é redução de custo computacional e aplicabilidade no mundo real.

2.3.1 Análise Baseada em Movimento

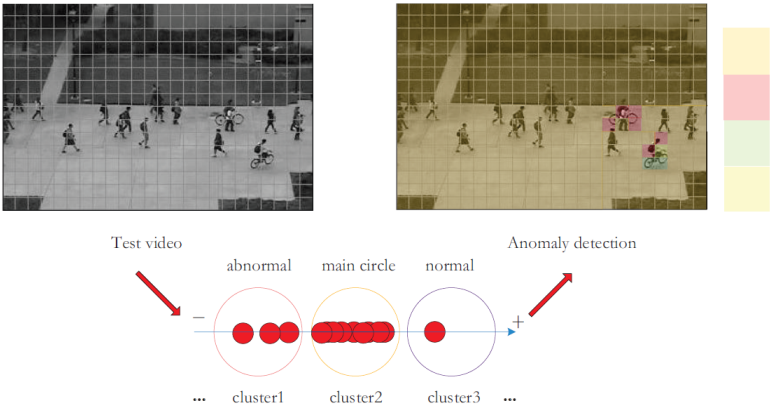
Os trabalhos baseados na extração do fundo estático da cena usam técnicas relacionadas com as alterações do *background* de cada *frame* ao longo da sequência de vídeo. Processar como um todo as mudanças de informações na evolução de *frames*, sem a necessidade de tratar um pipe-line prévio de processos como o *framework* sugerido da Figura 8, simplifica o processo da análise da cena pois não há etapas intermediárias que necessitem ser tratadas, exceto pela extração do *background* em alguns casos. No entanto a complexidade dos métodos aumenta. Exemplos de trabalhos nesta classe tem como base o fluxo ótico (HUANG et al., 2009), a distribuição de energia no espaço e tempo (ZAHARESCU; WILDES, 2010), PHD(Probability Hypothesis Density)(MAGGIO; CAVALLARO, 2009), PCA (YU; MOON, 2009; TZIAKOS et al., 2010), GMM treinado com EM em sub-regiões (MAHADEVAN et al., 2010) e mapas de direção (GRYN et al., 2005).

Uma abordagem interessante dessa classe foi a proposta por Zaharescu e Wildes (2010) onde a avaliação da distribuição de energia no espaço

e tempo são utilizadas para modelar comportamento. Segundo os resultados obtidos pelos autores, esta representação pode capturar uma grande variedade de padrões visuais que ocorrem naturalmente no espaço e tempo. Um modelo de distribuição de energia orientada espaço-temporal é gerado para modelar o comportamento e comparado com os novos movimentos observados. O método ainda atua mesmo quando somente um subconjunto do modelo é apresentado para a nova observação. A abordagem considerada singular nessa classe merece uma atenção em função de que ela somente concentra esforços computacionais somente nas regiões onde estão ocorrendo mudanças importantes de energia e desse modo, permitem o uso em cenas reais e tempo real.

Um outro exemplo desse tipo de análise é o modelo proposto por Guo et al. (2013) ilustrada na Figura 12. Para cada célula da grade eles usam um algoritmo que usa o deslocamento médio de intensidade de *pixels* (*mean shift algoritm*) e armazenam ali, via GMM todas as probabilidades encontradas. Após definição do limite mínimo de todas as probabilidades dessas células, qualquer célula em qualquer *frame* que apresentar uma probabilidade menor do que o limite estabelecido é então marcada como anormal.

Figura 12: Exemplo de análise baseada em região de um modelo com abordagem baseada em movimento.



Fonte: Modelo proposto por Guo et al. (2013).

Sudo et al. (2008) também apresentaram uma proposta que somente atua computacionalmente sobre o que ele chama de subespaços, regiões onde estão ocorrendo movimentos dos objetos. Eles realizaram uma modificação nos algoritmos SVM e PCA para torná-los com treinamento on-line não supervisionados incrementalmente. Segundo os autores a distância entre veto-

res capturados por PCA projetados dentro dos subespaços são aproximados com a distância no espaço original e por isso, depois de aprender por uma sequência longa de *frames*, o sistema consegue inferir sobre anormalidades através do classificador SVM de classe única.

Essas propostas são adequadas àquelas aplicações já citadas anteriormente como análise do movimento da forma de caminhar, dos gestos, de multidão, do corpo humano, etc. Nesses casos os movimentos são mais complexos e exigem mais processamento localizado nas regiões de interesse. Aplicações onde o contexto precisa ser levado em consideração, também se enquadram bem nessa linha de estudo.

Outro trabalho que ofereceu contribuições na linha de abordagem baseada em movimento aplicada a vários tipos de vídeos foi o de Saligrama e Chen (2012). Eles trabalharam com regras de decisão ótimas locais em sub regiões do *frame* para inferir sobre as dependências estatísticas espaciais e temporais globais dos movimentos.

A revisão bibliográfica se limitou em identificar somente algumas propostas e artigos que se encaixam na análise baseada em movimento sem se aprofundar nos métodos ou técnicas utilizados uma vez que esses trabalhos não foram os alvos para se discutir no presente trabalho.

2.3.2 Análise Baseada em Rastreamento

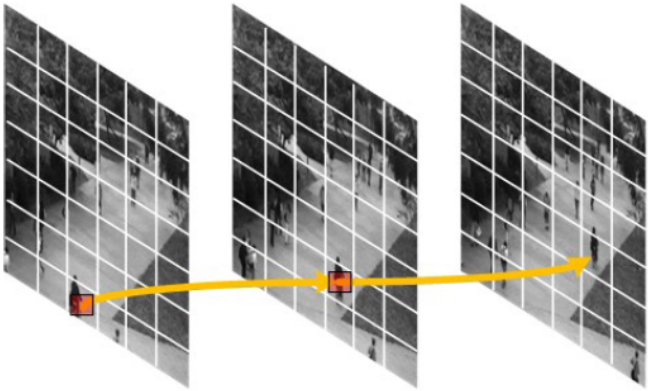
As abordagens baseadas em rastreamento necessitam das coordenadas 2D e das informações que discriminam cada objeto em cada *frame* ao longo do seu deslocamento nas sequências de vídeo. Esses dados serão computados por algum modelo estatístico que identifique padrões associados aos seus comportamentos. Em geral o rastreamento cria modelos durante uma fase de treinamento e após isso, na fase de observação, os desvios significativos encontrados são rotulados como anormais. O rastreamento é a única fonte de informação nessas abordagens e a confiabilidade dessa fonte para os mais diversos cenários continua sendo alvo de pesquisa em função de alguns desafios ainda em aberto nessa tarefa (CANNONS, 2008).

Diante disso, a qualidade ou a confiabilidade das inferências na análise de movimento fica dependente não somente do modelo adotado para este fim como também da robustez da fase de rastreamento que fornecerá os dados para análise. Isso leva as abordagens baseadas em rastreamento procurar estratégias que requerem menor custo computacional, a fim de tornar as aplicações que envolvam cenários do mundo real viáveis em diferentes contextos (SODEMANN et al., 2012). Os autores que se arriscam a desenhar propostas desde a captura dos *frames* de vídeo até inferência sobre o compor-

tamento dos objetos móveis, precisam determinar restrições de modo a tornar viável a grande carga de trabalho envolvida nas estratégias de cada parte do processo (XIANG; GONG, 2005; BERCLAZ et al., 2008; BASHARAT et al., 2008).

Zhang et al. (2013) apresentam uma elaborada proposta que determina anormalidades baseada na análise de trajetórias confrontadas com as predições realizadas em cada *frame*. A abordagem começa detectando regiões anormais baseadas na textura, tamanho e movimento dos objetos, as quais foram previamente treinadas com os vídeos sem anomalias. Após isso, os caminhos entre regiões construídos a partir das regiões anormais detectadas, chamados de *pathlets*, são usados para detectar as trajetórias anormais além de prever a localização dos objetos usando SVM. A Figura 13 ilustra a ideia onde nos dois primeiros *frames* é detectada localização com anomalias e é possível prever a localização da anomalia no terceiro *frame* adjascente.

Figura 13: Exemplo de análise baseada em região e de um modelo com abordagem baseada em rastreamento.



Fonte: Modelo proposto por (ZHANG et al., 2013).

Outras abordagens tratam sobre cenas do mundo real mas em geral, também limitadas em flexibilidades de cenários e alvos. Exemplo disso é o uso de técnicas como a lógica fuzzy como fez Hanapiah et al. (2010) onde heurísticas[2] são utilizadas para reduzir a complexidade da análise dos dados.

[2]Apesar da possibilidade de conduzir a erros, é comum encontrar o uso de heurística em alguma parte nos processos de análise de vídeo. Ela aparece sob a forma de condições de contorno ou de adoção empírica de limiares de decisão, baseada na experiência de observações similares, hipóteses ou intuições.

Nesse sentido, modelos estatísticos ganham um espaço importante para auxiliar na solução dos vários problemas que aparecem ao longo do processo. Reduzir a complexidade implica em reduzir o custo computacional ou complexidade dos algoritmos, determinantes na busca de performance e viabilidade das aplicações em tempo real.

2.3.2.1 Os Desafios do Rastreamento Robusto

Na prática o trabalho de rastreamento de objetos móveis frequentemente encontra mais outros problemas especialmente quando é necessário manter a identificação dos objetos sob longas durações de vídeo. Existem várias razões para isso nos mais variados contextos, principalmente nos cenários *outdoor*. Por exemplo, para os modelos de um rastreamento robusto visando rastrear pessoas é necessário manter sua identificação com as respectivas coordenadas no *frame* de vídeo controlando os seguintes obstáculos:

- Os fundos de cena real são complexos, variados e mutáveis;
- A variação de iluminação da cena ou movimentos indesejáveis ou fora do escopo de análise causados principalmente por condições climáticas;
- O rastreamento de múltiplos alvos que inclusive podem estar trocando de dimensões e ângulos de observação ao longo da trajetória;
- Identificar objetos não rígidos como uma pessoa e sua interação no ambiente devido a ambiguidade causada por articulação do corpo, roupas soltas, e oclusão entre as partes do corpo, entre corpos ou entre corpos e um objeto fixo do fundo estático da cena;
- Objetos que permanecem estáticos por uma sequência longa de *frames*, por exemplo, maior que 100 *frames*, podem acabar sendo interpretados como parte do fundo de cena.

Para outros tipo de objetos, além dos citados acima outros problemas podem emergir. No rastreamento de veículos de diversos tipos por exemplo, a variação das velocidades no deslocamento em relação aos outros, pode dificultar a manutenção da correspondência do seu traço de deslocamento. Observa-se que para conseguir uma robustez no rastreamento visando superar esses vários desafios cada técnica, método, modelo ou abordagem deve ser específica para cada situação desde os primeiros níveis de abstração dos sinais de vídeo para o sistema, passando pela sua segmentação até a sua análise em alto nível. O número de combinações dessas escolhas torna dinâmico esse

tipo de pesquisa permitindo que se possa conectar as melhores abordagens como boas opções para propor novas abordagens na expectativa de alcançar melhores resultados e maior amplitude de aplicações e portanto com liberdade de uso em diferentes contextos.

Para avaliar sobre o comportamento de um objeto rastreado, em geral adota-se um modelo de rastreamento que contorne um ou mais dos obstáculos citados anteriormente de acordo com o contexto da aplicação. A próxima etapa é aprender sobre o comportamento do trajeto de cada tipo objeto móvel reconhecido no ambiente. Neste ponto, ainda dependendo da aplicação em foco, outros desafios surgem para serem considerados e controlados:

- Identificar que um mesmo traço de deslocamento pode ser normal ou anormal de acordo com o contexto de diferentes ambientes;
- Aprender sem supervisão sobre o comportamento do movimento em longos períodos de observação dos cenários;
- Ao longo do tempo, movimentos anteriormente rotulados como anormais, podem assumir identificação contrária;
- As implementações das técnicas adotadas nas etapas anteriores somadas ao processo de análise de comportamento vão exigir uma grande capacidade de computação limitando muitas vezes sua aplicação em tempo-real.

No estado da arte sobre rastreamento, Cannons (2008) fez uma compilação ampla de ferramentas, abordagens, técnicas e modelos para demonstrar como diversos autores tem tratado um ou mais dos obstáculos citados no início desta seção. Ele e outros autores (XU et al., 2010; KO, 2008; HU et al., 2004) dividem estes modelos em três principais categorias: Rastreamento usando características discretas, rastreamento com contornos e rastreamento baseado em região, que serão explorados em síntese nos subitens a seguir.

i. Rastreamento Baseado em Características

No rastreamento usando características discretas o objeto móvel é rastreado a partir de um simples ponto ou em versões mais complexas as características rastreadas podem ser linhas, grupos de bordas ou até mesmo modelos 3D. Quando se usa um ponto como referência, sua posição e velocidade são as variáveis para corresponder cada alvo ao seu traço de deslocamento *frame* a *frame*. As propostas de Narayana e Haverkamp (2007), Basharat et al. (2008) e Zeng e Chen (2011) estão nessa linha de rastreadores.

Em geral, após a extração do *background* e a posterior segmentação dos objetos móveis, determina-se o centro médio de seus *pixels* ou **centroide** o qual será o ponto a ser rastreado na sequência de *frames*. O processo é determinístico e bastante carregado de heurísticas. Caso sejam usadas linhas ou bordas do objeto é necessário usar ferramentas de predição como filtros de Kalman para predizer e corrigir os parâmetros dessas linhas nos *frames* subsequentes.

Os desafios aumentam quando se deseja rastrear o melhor alinhamento de um modelo 3D do objeto com as medidas retiradas da imagem rastreada. Uma das abordagens também utilizadas nessa categoria de rastreador é o uso do método detector de cantos e bordas de Harris (HARRIS; STEPHENS, 1988). Em resumo, esse método detecta cantos em uma pequena área onde existe o encontro de bordas de um objeto que possuem diferentes direções na imagem captada como um plano 2D projetado na câmera. Esses pontos mantém relações em vários ângulos de observações e portanto caracterizam uma reta, plano ou planos importantes do objeto. Estes pontos passam a ser o foco no rastreamento.

Uma boa questão a ser resolvida nestes métodos é decidir quais serão os melhores pontos para escolher em um objeto que pode apresentar várias possibilidades de escolha. Após usar um detector de cantos como o método de Harris pode-se utilizar um método como o KLT (Kanade- Lucas-Tomasi Feature Tracker) (SHI; TOMASI, 1994; SUN; GUO, 2008) ou o método com SIFT (Scale Invariante Feature Transform) (YANG et al., 2009; HU et al., 2008). O método SIFT é um descritor que se destaca nesta tarefa porque ele é minimamente sensível as mudanças de escala, iluminação e aparência do objeto, características essas costumeiramente presentes durante um rastreamento (HU et al., 2008; MOREELS; PERONA, 2005). Este descritor é baseado na magnitude do gradiente e na orientação de todos os *pixels* em uma região em volta de um ponto chave ou *Key point*. Eles são colocados em uma janela gaussiana e acumulados em um histograma resumido em sub-regiões. A magnitude de cada orientação corresponde a soma dos vetores de mesma direção da região. A distancia entre os histogramas definida como **descritor SIFT** é usada como medida de correlação. O cálculo de distancia euclidiana pode ser usado para calcular a distancia entre os histogramas. Se a distancia for inferior a determinado limite é porque este é um ponto importante do objeto e deve ser utilizado como descritor. O SIFT potencializa a caracterização dos objetos, em especial os tridimensionais que em geral são os alvos de rastreamento em videovigilância. Uma de suas desvantagens é a complexidade computacional ($O(\cdot)$) quando implementado conceitualmente, consumindo

$O(n^2 \sim n^3)$ na etapa de treinamento e $O(n)$ no teste. No entanto autores como Yang et al. (2009) já apresentaram soluções usando representação de **códigos esparsos** dos descritores SIFT e reduziram a complexidade para $O(n)$ no treinamento e para uma constante na fase de teste.

O SIFT tem marcado presença em um número crescente de trabalhos relacionados com robustez na identificação de padrões especialmente pela sua insensibilidade aos vários obstáculos citados anteriormente. A identificação de um vetor estável com dimensão reduzida relacionado com diversos tipos de objetos 3D motiva o uso associado com classificadores robustos como o SVM.

Outras soluções como de Javed e Shah (2008) pode ser considerado como um rastreador dessa categoria. Os autores desenvolveram uma estratégia de criar um vetor de características chamado de Movimento Recorrente da Imagem (RMI) para calcular movimentos repetidos dentro da região do blob. Segundo eles, diferentes tipos de movimento dentro da área de seus blobs possuem diferentes tipos de RMI e então podem ser classificados em diferentes categorias. A técnica proposta por eles também é oportuna para detectar objetos esquecidos ou carregados.

ii. Rastreamento Baseado em Contorno

O rastreamento de contornos oferece um caminho alternativo de implementação de rastreadores em função de levar em conta a parametrização discreta de curvas abertas (*snake-based*) ou fechadas (*level set contour*) que envolvem os limites da forma do objeto rastreado. Esta categoria de rastreamento não é facilmente implementável pois exige uma modelagem baseada em um equacionamento que dependente das formas a serem rastreadas. Além disso pressupõe-se que as mudanças de forma entre um *frame* e outro não devem ser significativas para não se perder a identidade com a formulação correspondente. Isso implica que a captura dos *frames* deve ser rápido o bastante para preservar essa relação (em geral maior que 25 *frames* por segundo).

As primeiras propostas deste tipo de rastreamento envolveu os conceitos de energia que conseguem representar bem contrastes, cores e bordas. Outras propostas incluíram descritores regionais como cor, textura, histogramas, em função das limitações do uso de informações apenas ao longo do contorno em si. Novamente o papel de estimadores com o filtro de Kalman e PCA se revelam como usuais nessa categoria. Um exemplo desta abordagem pode ser vista em ZHANG et al. (2009). Os autores criam uma base de dados da silhueta dos objetos a partir de

um modelo de mistura de gaussianas adaptativo. A dimensionalidade dos dados é reduzida usando PCA e um algoritmo chamado de Marginal Fisher Analysis (MFA). O rastreamento segue então combinando as técnicas de PCA e MFA, e diferentes objetos de três grupos, pedestres, automóveis e outros, são reconhecidos e rastreados com muita robustez.

iii. As abordagens Baseadas em Região

Por fim, o rastreamento baseado em região que é uma coleção de rastreadores que usam técnicas estatísticas para inferir as próximas posições dos objetos no *frame* seguinte totalmente baseadas na informação contida na região de *pixels* do objeto destacado do fundo estático da cena. Essa informação pode ser cor, textura, gradiente, energia espaço-temporal, resposta de filtros ou combinação dessas. Os autores (CZYZEWSKI; DALKA, 2008) abordam uma proposta que considera a cor para discriminar o fundo estático da cena com os objetos móveis onde cada *pixel* da imagem é descrito por misturas de gaussianas dos componentes RGB de cor. O rastreamento segue usando filtros de Kalman com vetores de 6 e 8 dimensões que caracterizam cada objeto móvel na cena associado a uma matriz que relaciona cada traço de deslocamento com um blob.

Apesar da divisão de categorias ter um número pequeno, o número de combinações de técnicas e abordagens com diversas ferramentas é significativo. Em todos os tipos existem aplicações que mais se alinham com a técnica usada na sua concepção. Por exemplo, um rastreador que usa linhas do objeto como características a serem rastreadas em um próximo *frame*, não é apropriado para reconhecer objetos curvos ou circulares. Estas particularidades são adequadas para muitas aplicações no entanto são sensíveis ao contexto.

As três categorias descritas anteriormente refletem a maior parte de trabalhos propostos para atacar especialmente o problema de associação de dados no rastreamento. Observa-se que todas usam em comum um modelo de aparência do objeto para inferir corretamente sua nova posição em um próximo *frame*. Por isso pode-se enquadrar essas categorias em uma sub-classe de abordagens baseadas em modelos de aparência, em contraste com outra sub-classe baseada em em Mapas de Probabilidade de Ocupação (POM). Nessa linha encontra-se o trabalho de (BERCLAZ et al., 2008) que identifica e divide o plano de chão no FOV de múltiplas câmeras em subregiões que definem um mapa de ocupação de múltiplos objetos. Nesse caso, modelos de aparência dos objetos também são usados para calcular as probabilidades nas áreas do mapa e assim controlar completamente os problemas

de oclusão.

Independente do uso de qualquer estratégia para detectar movimento anormal em uma abordagem baseada em rastreamento, os dados que irão compor o conjunto de referência para análise serão originados pela própria categoria do rastreador escolhido. Assim, o modelo de movimento fornecerá os dados relativos a localização no espaço (região ou *pixel*) dos objetos rastreados e o modelo de aparência fornecerá os dados relativos a caracterização correspondente. Essa caracterização pode ser definida como uma única variável que representa o tipo de objeto (neste caso apoiado com um processo adicional de classificação), descritores simples como largura e altura do blob ou descritores mais completos como os *key points* de um SIFT ou histogramas locais, entre outros.

Obviamente que a quantidade de informação utilizada para caracterizar cada objeto vai impactar no trato computacional. Cabe então adotar modelos de movimento e de aparência que melhor se aderem ao propósito da análise de movimento de múltiplos objetos com múltiplas aparências.

2.3.3 O Conjunto de Vídeos de Referência - O Dataset

O modelo ideal de um sistema concebido para fazer a análise de cenas de vídeo seria aquele que consegue detectar anormalidades dos objetos presentes nas cenas sem qualquer treinamento ou conhecimento anterior do comportamento tanto no contexto quanto na movimentação usual. A pesquisa nesta área ainda está longe de alcançar resultados com este grau de autonomia ou cognição devido a complexidade de dados e suas relações no envolvimento desde a captura de vídeo até a análise.

Para contornar os desafios de um cenário ideal, a avaliação do desempenho dos algoritmos ou sistemas é focada na busca de melhores resultados sobre métricas específicas de um conjunto de vídeos de referência chamados de *datasets*. Muitos desses vídeos são disponibilizados de forma pública e explorados por grupos de pesquisa espalhados pelo mundo. A versão mais básica de um *dataset* é aquela que possui somente sequências de imagens em formatos usuais de extensão de arquivos *.jpeg* ou *.tiff* ou vídeo em formatos como *.avi*, *.mpg* ou *.mp4*. Os *datasets* em versões mais completas, acompanham além das sequências de imagens ou vídeo, anotações associadas em cada *frame* de vídeo através de dados ou metadados em vários formatos de arquivos. O enriquecimento de informações sobre um vídeo, além de garantir uma referência para a avaliação de algoritmos nas abordagens sobre sua análise, permite avaliar de forma isolada vários processos de um *framework*

como aqueles vistos na seção 2.2. Então, se existe o detalhamento adequado nas anotações de vídeo, é possível avaliar o desempenho de algoritmos específicos somente na fase final da análise, sem que seja necessário propor um modelo que trate todas as suas etapas iniciais ou intermediárias. Por conta disso, o uso da anotação de vídeo foi fundamental para o desenvolvimento deste trabalho, que estabeleceu como meta, atuar somente na análise de anormalidade de rastros já idealmente identificados.

Os *datasets* são sequencias de vídeo capturadas em períodos determinados a partir de câmeras instaladas em ambientes *indoor* ou *outdoor*. Estas cenas podem ser captadas sem adulteração do movimento natural dos objetos (cenas do mundo real) ou podem ser resultado de uma produção envolvendo atores e outros objetos em movimentos ou ações planejadas para fins de análises específicas. Outra opção ainda é o uso de simulação de cenas totalmente criadas de forma virtual através de softwares específicos para este fim. No caso de produção específica, os vídeos são segmentados em dois conjuntos de vídeos: O vídeo de treinamento e o vídeo de teste.

O vídeo de teste possui uma ou mais sequencias de *frames* daqueles comportamentos anormais que se deseja detectar ou analisar. Em geral os vídeos de teste são em menor número do que os vídeos de treinamento. As sequências de testes possuem de forma aleatória, a presença de movimentos não usuais que devem ser detectados pelos modelos.

As sequências de vídeo para o treinamento possuem conteúdos que não apresentam anomalias de movimentos dos objetos, visando permitir que os modelos criados reconheçam informações relativas aos comportamentos normais das cenas. Dependendo da abordagem do treinamento, o conjunto de *frames* para esta etapa do processo pode conter cenas com movimentos anormais e normais, ou seja, uma sequência de *frames* sem "cortes".

A medida de desempenho de um modelo para DMA baseado em rastreamento geralmente é realizada usando a quantidade de acertos nas inferências em relação a todos os eventos anotados no vídeo de teste. Ou seja, tomando-se como positivo o acerto de um evento anormal, a medida é melhor se o algoritmo avaliado infere corretamente tanto a detecção de um evento anormal (verdadeiro positivo) como também a detecção de um evento normal (verdadeiro negativo) enquanto eles acorrem. Uma variação dessa medida pode ser aquela em que se avalia o quão rápido ocorre a detecção do evento anormal, assim que ele começa a acontecer e até o momento em que ele termina. Nesse caso a medida de desempenho passa a ser a relação entre a quantidade de *frames* detectados como anormais em relação ao total de *frames* de cada evento anotados como tal. No caso de modelos de DMA baseada em região, outras medidas ainda podem ser adotadas como por exemplo a precisão da forma em *pixels* do objeto que segue uma trajetória anormal *frame* a

frame.

Os *datasets* constituídos por sequências únicas de vídeo sem a distinção entre teste e treinamento, são usados nos modelos de treinamento não supervisionado e por isso passam a ter importância nesses casos. Esses cenários são os mais interessantes para alcançar objetivos em aplicações visando o mundo real. Uma proposta nessa direção é encontrada no artigo de (XIANG; GONG, 2008) onde um conjunto de dados chamado de *bootstrapping* com alguns segmentos de vídeo aleatoriamente escolhidos em um número menor do que o total de amostras é usado para inicializar o modelo de uma DBN (Dinamic Bayesian Network). Os autores também adotaram um outro modelo chamado de Teste da Taxa de Probabilidade (LRT - Likelihood Ratio Test) para detectar comportamentos anormais de forma incremental e adaptativa a partir do treinamento inicial com o *bootstrapping*. Segundo os resultados desses autores, a modelagem proposta permitiu aprendizado totalmente não supervisionado incluindo trocas de contexto.

A demanda atual de *datasets* visa atender um número predominante de trabalhos para modelos supervisionados. Observa-se que este caminho é o mais conveniente para extrair e conhecer os melhores resultados de muitas abordagens em análise de movimento. Será uma tendência natural o crescimento de propostas voltadas para os modelos não supervisionados pois eles representam melhor o mundo real no qual se aprende pela experiência. A inteligência artificial nesses casos cumpre um papel fundamental como parte da modelagem. Algumas das conferências e workshops citados na introdução deste capítulo lançam ou adotam *datasets* como base e desafio para autores candidatos na submissão de artigos. A partir dos resultados iniciais outros trabalhos são desenvolvidos, motivados pelo aprimoramento de soluções. É o caso dos *datasets* produzidos pelo Statistical Visual Computing Laboratory (SVCL) da UCSD - University of California, San Diego. Na forma mais comum, os *datasets* são concebidos e disponibilizados como referencias para o desenvolvimento de trabalhos relacionados em diversos congressos ou workshops. Exemplos deles são as várias versões do PETS (Performance Evaluation of Tracking and Surveillance), o CAVIAR (Context Aware Vision using Image-based Active Recognition), o VIRAT, dedicado para pesquisas em Continuous Visual Event Recognition, o qual incorpora uma série de vídeos de longa duração de cenas realísticas, diferenciados com maiores resoluções e captados de várias câmeras fixas ou aéreas de vários tipos de eventos (OH et al., 2011), o BEHAVE (computer-assisted prescreening of video streams for unusual activities), entre outros.

Outros *datasets* são virtualmente construídos por aplicativos de softwares visando atender fins específicos tais como modelos matemáticos na análise de multidão (ou análise de movimento denso de objetos, crowd analysis).

Diante da raridade da captura de cenas reais para esse tipo de análise, a virtualização é um recurso providencial pois é possível simular inúmeros cenários, dispensando a criação de uma produção encenada por atores, voluntários ou figurantes. Ferramentas de autoria desses tipos de cenário podem incluir o uso de agentes inteligentes fazendo com que cada objeto virtual se comporte de modo programado diante de diversas situações. Exemplos do uso desses *datasets* são citados no trabalho de Jacques Junior et al. (2010). Embora se construam ambientes artificiais totalmente controlados, também é possível simular situações próprias de cenários reais como variações de luminosidade, fatores climáticos, aleatoriedade de movimentos, programação própria de cada objeto como um agente inteligente, entre outras.

2.3.4 Anotação de Vídeo

Em uma avaliação geral, observa-se que existe um esforço para se criar referências universais de *datasets* para conduzir a pesquisa nos diferentes ramos da visão computacional. No entanto, diante da infinidade de propósitos, ainda não há padrões de fato que regem a construção de conjuntos de dados de referência. Nesse sentido, os *datasets* mais presentes nos trabalhos relacionados com o tema desta tese, são aqueles que contém anotações de vídeo úteis de modo que possam ser usadas como métricas para avaliação dos algoritmos ou modelos propostos.

Anotar um vídeo significa agregar à ele informações relevantes e computacionalmente tratáveis no sentido de permitir ou dar suporte na construção de algoritmos para análise correspondente. Exemplos de anotações podem ser:

- **de metadados do vídeo**: são dados de referência que descrevem dados gerais da sequência de vídeo como tipo ou contexto do vídeo, quantidade de objetos ou rastros normais ou anormais, geolocalização, tipo de objetos, dados de calibração, entre outros;
- **de dados do vídeo**: identificação do inicio e final de *frames* que contém anormalidades;
- **de dados do *frame***: identificação de objetos que descrevem um comportamento anormal;
- **de dados do objeto**: identificação do tipo de objeto, forma, altura ou largura (*bounding box*), textura, histograma local, entre outros;
- **de dados da ação do objeto**: identificação do estado ou comportamento do objeto como saltando, deitando, correndo, movimentando

membros, expressões da face, entre outros;

- **de dados de localização do objeto** ou partes deste: identificação das coordenadas no plano ou das três dimensões, calibradas ou não, ângulos de movimento, coordenadas de membros, entre outros.

Os dados ou os metadados anotados devem traduzir ou reproduzir com fidelidade os comportamentos dos alvos captados pelas câmeras pois eles serão as referências para o treinamento e testes dos modelos de análise de vídeo. As anotações não possuem um padrão de organização ou de quantidade de dados pois isso vai depender do objetivo da modelagem e das ferramentas computacionais adotadas para implementação de cada proposta.

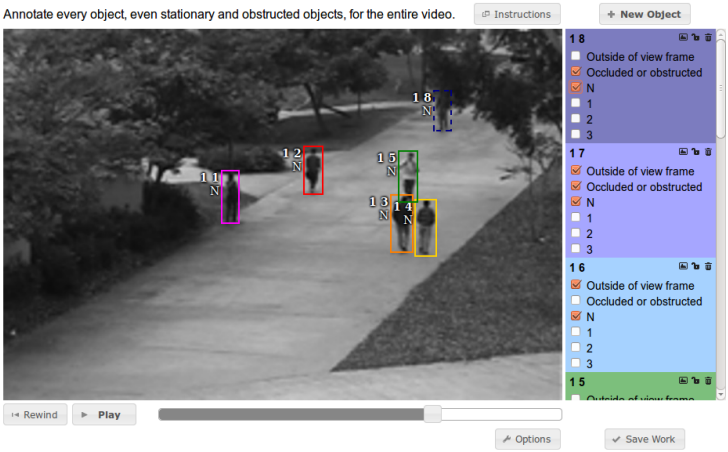
As anotações de vídeo em geral são representadas por arquivos no formato de texto (ex.: *.txt*) ou de linguagens de marcação (ex.: *.xml*), mas podem também estar disponíveis em outros formatos como o de planilhas eletrônicas (ex.: *.xls, .odt*) ou de ferramentas como o MATLAB - MATrix LABoratory. Assim, o conteúdo desses arquivos é um conjunto organizado e pré-definido de dados relativos ao vídeo e/ou de cada *frame*, os quais são registrados por alguma ferramenta de autoria ou edição de vídeo.

Como não existe uma ferramenta automática para anotação de vídeo, quanto mais dados pretende-se anotar, mais tempo e preciosismo é necessário dispensar para esta tarefa. O trabalho se torna imensamente maior se o número de objetos de interesse e a quantidade de *frames* é maior. Essa é uma das razões pela qual é mais raro encontrar trabalhos voltados para análise de vídeo baseada em rastreamento de objetos e/ou vídeos de longa duração.

Grande parte dos *datasets* dispõe de anotações de vídeo limitadas pelos objetivos específicos dos grupos de pesquisa, os quais desenvolvem ferramental próprio para esta tarefa. Vivenciando esse problema, surgiram algumas iniciativas de autores que criaram e disponibilizaram tais ferramentas como o ViPER (The Video Performance Evaluation Resource) (University of Maryland, 2005) o LabelMe de Yuen et al. (2009) e a proposta de Kavasidis et al. (2014). Dentre essas e outras ferramentas pesquisadas, foi encontrado a VATIC - Video Annotation Tool from Irvine, California desenvolvida por Vondrick et al. (2013). O trabalho desses autores foi além da concepção de uma ferramenta que tornasse a tarefa mais automatizada. Eles criaram um ambiente colaborativo baseado na computação em nuvem com uma linguagem e procedimentos próprios que permitem a adesão de quaisquer candidatos interessados em realizar a tarefa de anotação. Como contrapartida, o trabalho é remunerado de acordo com tipo de vídeo, tempo dedicado e outros quesitos. A qualidade da anotação é garantida por um treinamento prévio da ferramenta e a devida aprovação de quem contrata o serviço. O VATIC ainda tem a possibilidade de execução em modo *offline* permitindo assim, que a

anotação possa ser realizada desconectada da nuvem. A Figura 14 ilustra a tela do VATIC editando um *frame* de vídeo já anotado. Diante das características encontradas na ferramenta, ela é adequada para *datasets* usados em trabalhos que não dependem da forma dos objetos. A quantidade elevada de objetos móveis também não é uma restrição, embora a anotação seja mais lenta e meticulosa.

Figura 14: Amostra de um *frame* com anotação de vídeo.



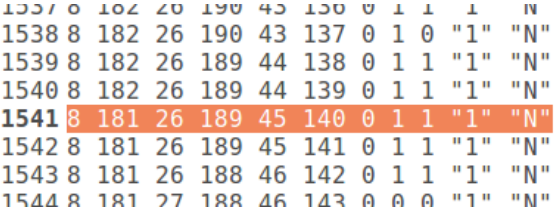
Fonte: Anotação de um *frame* de vídeo realizada pelo autor no *dataset* da UCSD (MAHADEVAN et al., 2010) usando a ferramenta de anotação de vídeo VATIC (VONDRICK et al., 2013).

A disposição das informações na amostra da Figura 14 são próprias do VATIC. Na área do *frame* pode-se observar seis pessoas que estão se deslocando em sentidos diferentes. O número mais à esquerda "1" no canto superior esquerdo do *bounding box* identifica o tipo de objeto seguido logo a direita por outro número sequencial de objetos de mesmo tipo. Logo abaixo há um outro algarismo ("N") que rotula que tipo de movimento o objeto em questão está descrevendo. Neste exemplo, o *frame* faz parte de um vídeo de treinamento e os seis objetos descrevem uma trajetória normal conforme o significado do algarismo pré-configurado como "N".

Outro dado contido nesta anotação é a condição de oclusão do objeto tipo "1" de sequencial "8" que está representado pela linha tracejada em seu *bounding box*. A síntese das informações desse objeto sequencial "8", no *frame* de número "140", ocupa a linha de número "1541" de um arquivo texto, a qual pode ser vista destacada na Figura 15. Os demais dados representam

coordenadas 2D, largura e altura do *bounding box* e condições do objeto no *frame* (visível, ocluso e tipo de anotação).

Figura 15: Amostra de uma sequência de vetores de anotação de vídeo.



Fonte: Arquivo gerado pela execução de comando do aplicativo VATIC.

Os resultados gerados por uma ferramenta como a VATIC pode ser útil não só para a etapa final na análise de vídeo como também para as etapas intermediárias que envolvem o rastreamento e classificação de objetos e segmentação de movimento. Com as anotações é possível isolar uma parte de um *framework* e avaliar especificamente modelos propostos neste ponto, assim como foi feito no presente trabalho.

2.4 DETERMINAÇÃO DO TAMANHO IDEAL DE REGIÃO NOS MODELOS BASEADOS EM REGIÃO

Trabalhos anteriores de outros autores, usaram a subdivisão do ROI na cena, a fim de tornar o processamento mais eficiente ou controlável, bem como métodos para reduzir a dimensionalidade e custo computacional (ELHOSEINY et al., 2013). A chamada *maldição da dimensionalidade* avaliada por Bishop (2006), é um tema recorrente, que requer uma abordagem mais sofisticada em dados n-dimensionais quando n é maior do que 3.

A redução de dimensionalidade pode ser realizada tanto pela seleção de um subconjunto do espaço de características de um modelo de aparências quanto por técnicas supervisionadas ou não de transformação dessas características. Os autores Tziakos et al. (2010) utilizaram uma grade de regiões em um detector de movimento anormal local para testar os efeitos da redução de dimensionalidade. Mesmo tendo técnicas como PCA e SVM para implementar a redução do espaço dimensional, trabalhos como os de Zhang et al. (2013) e Saligrama e Chen (2012), usam dividir a ROI ou o *frame* em regiões menores e ali aplicam suas estratégias. Esses autores concordam que é necessário o uso de técnicas baseadas em região, caso contrário é impraticável

aplicar muitas estratégias e modelos no mundo real.

Adotou-se no presente trabalho um método de modelagem de cena semelhante da estrutura implementada por Li et al. (2012a). No entanto, eles determinaram empiricamente o tamanho da região da grade. Como no trabalho de Feizi et al. (2012), o número de *pixels* no tamanho do conjunto é também convenientemente por eles determinado. Por outro lado, Kwon et al. (2013) usou o conceito de entropia para ajustar o tamanho da região, a qual deram o nome de *célula*, a fim de ajustar nelas, as melhores matrizes de dados que detectam movimentos anormais.

A abordagem do presente trabalho corrobora com o mesmo entendimento de Kwon et al. (2013), quando os autores afirmam que os pequenos deslocamentos de centroides do objeto em torno de uma vizinhança de *pixels* tem influência desprezível sobre a avaliação de movimento. Assim, ignorar a porção de dados que representam esses pequenos movimentos entre *pixels* vizinhos, não afeta significativamente a conclusão geral sobre a anormalidade de movimento.

Outras propostas continuam na estratégia da divisão do *frame* ou da ROI em sub-regiões mas nem sempre justificam claramente a definição do tamanho do agrupamento de *pixels*, grade ou de blocos (LIU et al., 2010; FEIZI et al., 2012; ELHOSEINY et al., 2013). Alguns autores justificam o tamanho dessas sub-regiões baseados em heurísticas apropriadas para cada cenário. Isso ocorre em (BERCLAZ et al., 2008) que discretiza em $30x45$ locações possíveis o plano do solo onde pessoas são rastreadas pois consideram que uma área de $20cm^2$ é o espaço mínimo ocupado por elas. Já Li et al. (2012a) e Zhang et al. (2013) adotam empiricamente a divisão de um *frame* em uma grade fixa e uniforme de $16x16$ regiões. Os autores Saligrama e Chen (2012) escolhem o tamanho de cada bloco baseado em uma dependência da quantidade de objetos, não detalhada, de modo que não interfiram um sobre os outros. Os tamanhos de blocos são definidos convenientemente para cada *dataset* avaliado inclusive de forma não quadrática como $30x20pixels$ e $120x240pixels$.

As razões particulares que levam os autores a definirem o tamanho da região a ser utilizado em cada trabalho, sugerem que não necessariamente elas possuem um tamanho ideal. Isso significa que é possível encontrar dentro de cada modelo adotado, tamanhos de região diferentes que poderiam conferir um melhor desempenho do que aqueles apresentados. Essa interrogação encontrada em diversas referências bibliográficas, levou o presente trabalho a propor em sua metodologia, uma forma de encontrar e revelar com segurança, o tamanho ideal da região para extrair o melhor desempenho do modelo implementado.

2.5 INDEPENDÊNCIA DE CONTEXTO

O contexto oferece uma direção alternativa na solução daqueles problemas da visão computacional pois podem usar o cruzamento de informações de maior nível de abstração do ambiente para dentro dos modelos. Em videovigilância por exemplo, o contexto ganha muita importância quando o objetivo é a detecção de **comportamento** anormal. Em geral, o uso de técnicas de mineração de dados são propostas para descobrir e adicionar regras usuais no espaço e tempo relacionadas com os movimentos normais de objetos na cena. Desvios dessas regras são encarados como anomalias. Jiang et al. (2011) sugere propostas como essa.

O uso do contexto exige uma segmentação, em geral manual, da área da cena e por isso se pode concentrar os esforços computacionais nessas regiões que irão inferir sobre o comportamento individual ou em grupos de objetos móveis. Li et al. (2012b) propõe por exemplo um *framework* que automaticamente aprende semânticas de comportamentos de contextos espaciais, de contextos temporais e da correlação de contextos para depois detectar anormalidades nas cenas. O método usa agrupamentos gaussianos para segmentar semanticamente regiões da cena baseado em um vetor de características de 10 dimensões de cada objeto móvel presente na cena.

Na mesma linha de soluções que aprendem o contexto das cenas para oferecer maior robustez na identificação de comportamentos anormais em cenários com múltiplos objetos, é o trabalho de Maggio e Cavallaro (2009) onde ao longo da evolução das cenas, a distribuição espacial de pontos de origem dos alvos e de eventos desordenados são incrementalmente aprendidos. Eles usam um filtro como o PHD (Probability Hypothesis Density) que identifica o centro de maior probabilidade de um agrupamento de pontos feito com GMM e segundo resultados dos autores, cada objeto ganha uma melhora no desempenho de seu rastreamento em função do contexto aprendido da cena (CZYZEWSKI; DALKA, 2008).

Na abordagem que está sendo proposta, também deseja-se avaliar a anomalia do movimento de múltiplos objetos móveis a partir de modelos probabilísticos de espaço e tempo. A anomalia de cada movimento vai estar intimamente ligada ao tipo de objeto que é representado com descritores que os distinguem dos demais. Ou seja, após a longa observação e consequente treinamento do sistema sobre cada tipo de objeto que entra e se movimenta na cena, espera-se aprender sobre o que é normal de sua participação. Qualquer desvio do movimento usual resulta em divergências da probabilidade prevista e anormalidades são identificadas. Estando a dependência do modelo ligada especialmente às características do objeto, o restante do cenário e dos outros diversos tipos de objetos móveis são variáveis que não vão influenciar

de modo significativo sobre a identificação de anormalidades de movimento sobre os outros. Como exemplo, uma pessoa pilotando uma motocicleta que se desloca em uma trajetória e velocidade usual de pedestres, só deve ser identificada como um objeto com movimento anormal se ela estiver sobre uma calçada. O sistema não sabe que naquela área de *pixels* do *frame* existe uma calçada, mas ele deve aprender que naquela área não é usual observar a trajetória de um objeto que possui descritores bem diferentes do que os de pedestres.

Segundo Morris e Trivedi (2008) o contexto, ou seja, uma quantidade maior de informação do domínio de conhecimento, é necessário quando se deseja analisar comportamentos complexos. Esse tipo de análise não é o propósito do presente trabalho. Nesse relacionamento, pressupõe-se que se o objetivo é identificar normalidades ou não do movimento de objetos, somente é necessário obter e conhecer as informações usuais sobre seu movimento. Desse modo podemos entender que estamos tratando de uma abordagem que tende a possuir um certo grau de liberdade de uso em diferentes contextos. Esse é um dos motivos que se buscou testar o modelo proposto aqui com vários *datasets* de referência e em cenas do mundo real, comparando assim seu desempenho com abordagens da mesma linha.

Embora o estado da arte neste tema agregue um número bem maior de trabalhos relacionados, procurou-se manter o foco associado principalmente para aquela linha voltada a abordagens baseadas em rastreamento as quais são particularmente mais adequadas para a análise de vídeo em longos tempos de duração. No entanto, como já mencionado, o rastreamento é uma tarefa ainda carregada de desafios devido a necessidade de resolver oclusões e associações de forma correta na análise com múltiplos objetos. Isso tem levado os autores a dirigirem mais esforços para abordagens baseadas em movimento mesmo diante das cargas computacionais envolvidas em suas soluções. O presente trabalho busca resgatar o estímulo à pesquisa estratégias baseadas em rastreamento uma vez que a DMA proposta, que começa exatamente quando a missão do rastreador termina, mostra um custo computacional tão reduzido na análise do movimento que é possível destinar grande parte dele para atender os processos antecessores. Nessa direção, o próximo capítulo explora como foram panejadas e modeladas todas as etapas da presente abordagem.

3 IMPLEMENTAÇÃO DA ABORDAGEM

A implementação de um *framework* que obedeça a uma abordagem baseada em rastreamento como aquele modelo genérico mostrado na Figura 8 da seção 2.2, necessitaria um trabalho intenso e dedicado a cada etapa. O desempenho da tarefa da detecção de anormalidades no final deste processo fica portanto, totalmente dependente do desempenho de cada etapa ao longo da cadeia.

Foi objetivo do capítulo anterior destacar que ainda existem vários desafios abertos que distanciam do ideal, os resultados da saída de cada etapa do *framework*. No entanto isso tem sido o motor da pesquisa nessa área.

Um ponto importante da abordagem baseada em rastreamento é a robustez desejada na saída do rastreador de objetos. Ou seja, dados que reproduzam com fidelidade tanto o movimento quanto a identificação do próprio objeto durante todo o deslocamento dele na cena. É a partir da saída do rastreador que efetivamente se inicia a análise do comportamento de cada movimento. Então, os dados esperados neste ponto são aqueles que caracterizam os objetos e respectivos movimentos, como os seus descritores e as informações dos trajetos correspondentes. A DMA faz parte de uma dentre outros tipos de análises que podem ser realizadas a partir dessas informações. Cabe então definir quais informações são essenciais para serem utilizadas na modelagem desta etapa.

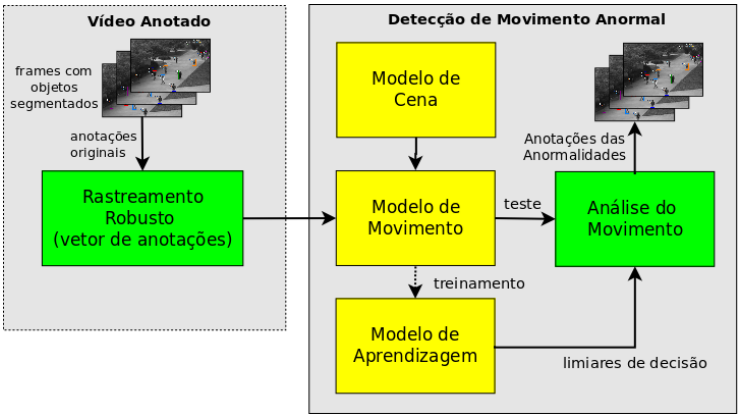
O foco deste trabalho foi direcionado somente na última etapa do *framework*, considerando que essa recebe como informações de entrada, os dados de um rastreador mais próximo possível do robusto ou do ideal. Dessa forma o desenvolvimento do presente trabalho se apoia sobre uma base de dados construída a partir de ferramentas de anotação de vídeo que, além de simular dados da saída de um rastreador, oferecem dados ideais para análise de movimento, isentos de incertezas ou imprecisões propagadas ao longo de todas as etapas previstas para um *framework* elaborado para este fim. Então, cada registro anotado em vídeo da transição de qualquer objeto móvel na ROI, produz um **vetor de anotação** com variáveis que caracterizam o objeto e seu respectivo movimento. Essa simulação compreende a substituição de todas as etapas anteriores previstas de um *framework* completo como aquele proposto por Ko (2008), discutido no capítulo 2.

Visando modelar e avaliar os resultados da implementação da abordagem proposta no presente trabalho, foram utilizados como referência para a simulação da saída de um rastreador robusto, vetores de anotação de dois grupos de vídeo: o primeiro utilizou vídeos anotados de um *dataset* nos quais foram realizados filtragem de trajetos e anotações complementares com ferra-

menta implementada pelo próprio autor e o segundo utilizou vídeos de outro *dataset* nos quais foram realizadas anotações fieis e completas através da ferramenta VATIC, também feitas pelo próprio autor.

A Figura 16 mostra as etapas que compõe a modelagem na DMA implementada neste trabalho as quais serão descritas com mais detalhes ao longo deste capítulo.

Figura 16: Modelagem da etapa de DMA.



Fonte: Elaborado pelo próprio autor.

Tomando como base o estado da arte discutido no capítulo anterior, a modelagem adotada para detectar trajetos anormais foi idealizada sob o fundamento de que, usando um modelo estatístico para este objetivo, é possível determinar um único valor de referência de probabilidade, ou limiar, que consiga separar corretamente o maior número de trajetos anormais dos normais. O limiar que otimiza a correta inferência nas duas classes possíveis (normais e anormais) vai depender completamente da modelagem e ferramentas adotadas. Mesmo com limiar particularmente definido como adequado em uma determinada abordagem, a quantidade de erros nas inferências ainda pode ser significativa.

Outro aspecto que influenciou a adoção da modelagem implementada neste trabalho, foi a busca de um equilíbrio entre o melhor desempenho nas inferências e a menor quantidade de informação a ser tratada nessa automação. Neste aspecto vale destacar que, mesmo considerando somente a última etapa em uma análise de vídeo, o tratamento e o treinamento dos dados gerados por uma sequência contínua de *frames*, requer a maior parte dos recursos e esforços computacionais em uma abordagem baseada em ras-

treamento. Minimizar estes custos significa viabilizar muitas abordagens a operar em aplicações do mundo real.

Sendo assim, o presente trabalho analisou os efeitos na movimentação de objetos entre regiões maiores do que um único *pixel* da resolução do vídeo visando comprovar o que foi mencionado no capítulo anterior, onde afirma-se que pequenos movimentos entre *pixels* vizinhos não alteram significativamente a conclusão sobre a anormalidade do movimento. Diante disso, o descarte da porção de dados que representa essa pequena movimentação, acelera o processo de treinamento e mantém praticamente inalterado o desempenho do modelo de reconhecimento de padrões adotado.

Com essa intenção, o detector de movimentos anormais mostrado na Figura 16 foi implementado sob o suporte de três modelos que atuam sobre os dados das anotações de vídeo: O modelo de cena, o modelo de movimento e o modelo de aprendizagem. Em linhas gerais, eles são brevemente explicados a seguir e serão detalhados nas próximas seções.

No **modelo de cena**, agrupa-se em cada região pré-definida na ROI, informações associadas a cada objeto que tem seu centroide pertencente a ela durante a observação em cada *frame* de vídeo. Essa região foi definida como uma área quadrada ou agrupamento com lado medindo p_u *pixels*, uniformemente arranjadas dentro de uma grade de regiões sobre a ROI, de acordo com as estratégias definidas aqui como **grade fixa** e **grade móvel** que serão melhor detalhadas na seção 3.2 deste capítulo. Com isso, constrói-se uma grade uniforme de regiões que pode possuir uma resolução igual ou inferior a resolução do *frame* de vídeo. Portanto, a quantidade de regiões na cena fica condicionada ao tamanho da ROI e ao valor de p_u, o qual fica aqui definido como **fator de grade**.

Nos dois tipos de grade, as regiões são numeradas sequencialmente da esquerda para direita e de cima para baixo, transformando a ROI do *frame* em um vetor unidimensional de regiões o qual foi definido como **vetor de regiões**. Neste aspecto, as 2 dimensões que representam a posição do centroide de um objeto são reduzidas para um escalar que representa uma posição no vetor de regiões. Portanto, para o caso de uma modelagem de cena com grade fixa, um fator de grade unitário ($p_u = 1$) representa o total original de posições de *pixel* do *frame*.

O **modelo de movimento** foi conduzido pelo foco da redução de custo computacional. Desse modo, a construção deste modelo se guiou sobre a meta da manipulação do menor número de dados possível. Para tanto, o modelo usa um vetor de dados, aqui denominado como **vetor de transição**, que usa os dados do vetor de anotação para formar um vetor com somente 3 variáveis (ou dimensões). São elas: o número da região (posição no vetor da grade de regiões), o tipo de objeto e seu *timestamp* no *frame*.

Os vetores de transição produzidos são usados na fase de treinamento do modelo de aprendizagem e depois, em um segundo momento, na fase de testes onde cada transição de objeto em sua trajetória é avaliada quanto a sua normalidade, apoiada pelos limiares de decisão aprendidos pelo modelo de aprendizagem. Os objetos que contém anormalidades em suas trajetórias são destacados dos demais durante a reprodução dos *frames* de teste a partir do momento em que eles desenvolvem um movimento anormal.

Por fim, o **modelo de aprendizagem** usou uma estratégia de treinamento supervisonada adotando GMM de cada região de grade parametrizada através do algoritmo EM. Para tanto foram utilizados, para cada sequência de vídeo, conjuntos de vetores de transição relacionados aos dados das anotações de *frames* que contém somente trajetos normais. São vetores resultantes do rastreamento do centroide de múltiplos objetos móveis. As rodadas de treinamento ocorrem em modo *off-line* e com classificação binária via ROC, para extrair de forma mais simplificada e eficiente, as referências dos limiares de probabilidade para serem usadas nas inferências sobre as anomalias do movimento em cada cenário. O modelo permite que seja disparada uma nova rodada de treinamento para encontrar um novo limiar caso existam mudanças significativas na ROI do cenário.

Além de assumir o uso de anotação de vídeo em substituição a cadeia de processos que antecede a análise de movimento, outras premissas foram determinadas para alinhar os resultados deste trabalho com seus objetivos. Na seção seguinte, elas são apresentadas.

3.1 PREMISSAS PARA CONSTRUÇÃO DO MODELO

Conforme observado na seção anterior, a DMA é a última etapa de um *framework* desenhado para uma abordagem baseada em rastreamento e portanto pode ser um processo com tratamento isolado dos demais. Conforme mencionado, para alimentar e avaliar esse processo, assumiu-se o uso de *datasets* com cenas reais outdoor de múltiplos objetos móveis. As anotações das sequências de vídeo escolhidas receberam manualmente a robustez de um rastreamento selecionando para isso, os melhores trajetos que envolvem predominantemente pessoas e veículos.

As anotações de vídeo, como dados básicos de entrada são amostras no formato de vetores de anotação 7-dimensional que representam informações do trajeto de cada objeto móvel. Cada amostra contém: as coordenadas 2D na resolução de *pixels* da cena, a largura e altura de *bounding box*, o *timestamp* da transição, o tipo de objeto e o tipo de movimento em curso.

O método proposto foi implementado em MATLAB®usando um com-

putador com processador Pentium Intel®Core™i5 CPU M450 @2.40GHzx4 com 6GB de memória RAM e sistema operacional UBUNTU 12.04 (precise) 64-bit. A medida do custo computacional, além de depender dos recursos computacionais utilizados, é sensível a estrutura dos algoritmos adotados na implementação do modelo, especialmente na fase de treinamento. Cada amostra vai exigir um lapso de tempo no escalonamento de processos realizados por um processador. Portanto, já que existe uma relação de proporcionalidade entre o custo computacional com o número de amostras envolvidas nos processos, é factível utilizar o total de amostras como métrica para avaliar resultados ou desempenho de algoritmos dentro de uma abordagem.

As subseções seguintes destacam outras considerações que guiaram toda a abordagem e as respectivas implementações.

3.1.1 Modelos de Referência

Os modelos de cena e de movimento utilizados aqui tomaram como ponto de partida os modelos propostos por Basharat et al. (2008) aplicando no entanto, uma estratégia particular de análise baseada em região maior que um único *pixel* do *frame*.

Esses autores desenvolveram um método para identificar anomalias tanto em movimentos locais quanto globais [1] usando um conjunto de vetores de transição armazenados em cada posição de *pixel* para modelar um GMM. Cada vetor é uma variável aleatória 5-dimensional que representa a próxima transição de cada objeto que passa por aquela posição. Assim, em cada *pixel* são armazenados inúmeros vetores de transição que representam as próximas transições de vários objetos que passam no mesmo *pixel*. Como contribuições adicionais, segundo os autores, as informações contidas nas variáveis aleatórias, após o treinamento com o algoritmo EM, permitiram criar funções de distribuição de probabilidade *pdf* que ajudaram a melhorar o desempenho da detecção de objetos através de uma realimentação de informação para a primeira etapa do *framework* proposto por eles. Essa informação consiste de valores médios calculados para cada *pixel* do tamanho mínimo de objeto e da taxa de aprendizagem, que são parâmetros fundamentais para a tarefa de extração do fundo estático da cena.

O trabalho de Basharat et al. (2008) se dedicou propor uma solução completa para uma abordagem baseada em rastreamento, no entanto observa-se o grande esforço computacional necessário para realizar o treinamento de

[1] Segundo os autores, os movimentos locais são aqueles analisados na transição imediatamente posterior à posição atual do objeto móvel, sendo as transições seguintes associadas ao movimento global.

muitas amostras para produzir uma *pdf* representativa para cada *pixel*. Isso fica evidente quando os autores propõe multiplicar as mesmas amostras em toda a região do *bounding box* de cada objeto e para cada transição. O recurso torna mais densa a quantidade de amostras por *pixel* para viabilizar o treinamento do GMM sem que se necessite adicionar mais trajetos nessa fase, todavia multiplica demasiadamente as mesmas informações.

O trabalho aqui proposto vai no caminho inverso do recurso utilizado de multiplicação de amostras. Ao invés de reproduzi-las na vizinhança, amplia-se a região além de um *pixel* para capturar amostras vizinhas e somente a partir daí, realiza-se a criação de *pdf*s. Os efeitos dessa estratégia serão discutidos no próximo capítulo.

3.1.2 Datasets de Referência

Foram utilizados dois *datasets* bastante distintos em relação as anotações de vídeo: O Ped2 da UCSD e três sequências contendo de $1 \sim 4$ horas de vídeo do projeto LOST [2] (Longterm Observation of Scenes with Tracks Dataset) disponibilizados pelos autores Abrams et al. (2012).

O Ped2 da UCSD compõe-se de 16 sequências de vídeo que totalizam 2550 *frames* sem a presença de movimentos anormais, destinadas para o treinamento do modelo e outras 12 sequências, diferentes da primeira, totalizando 2010 *frames*, mas com várias situações de movimentos anormais, destinadas para a fase de testes. Os movimentos anormais estão associados a presença de objetos não usuais como bicicletas, automóveis, ou de trajetos anormais como parada, velocidades diferentes ou cruzamento por áreas não usuais. Embora seja um *dataset* de um vídeo de muito curta duração, em torno de 2,5 minutos no total, e elaborado para fins de pesquisa em análise de multidão (crowd analysis), e portanto apropriado para abordagens baseadas em movimento, o Ped2 representou o equivalente a quantidade de anotações iniciais comparáveis as dos vídeos do LOST. Esse foi um dos fortes motivos que levaram a adoção da VATIC como ferramenta para gerar anotações que simulam com fidelidade a saída de um rastreador robusto, para diversos cenários e contextos. Só as automaticidades como as proporcionadas pelo VATIC, permitem anotar com relativa facilidade vídeos com muita densidade de objetos móveis, e inclusive multidão.

O projeto LOST compreende a disponibilização de vários *datasets* de vídeos automaticamente anotados construídos a partir do *streaming* de *webcams* outdoor, capturados e organizadas por números (de 1 a 25) na mesma meia hora todos os dias, em vários locais do mundo. A coleta de dados

[2] Disponível em *http://lost.cse.wustl.edu*, acesso em 04/08/2014.

começou em junho de 2010 e continua até fevereiro de 2015. Os *datasets* contém metadados da geolocalização, detecção de objetos e os respectivos resultados de rastreamento de diversos objetos móveis ao longo de diferentes resoluções de vídeo. Esse *dataset*, dentre outros avaliados como dos autores Oh et al (OH et al., 2011), veio de encontro aos objetivos deste trabalho, principalmente porque fornece anotações em vídeo do rastreamento de diversos tipos de objetos em contextos diferentes.

Os algoritmos desenvolvidos pelos autores que constroem automaticamente as anotações de vídeo, não possuem compromisso com a fidelidade em função dos objetivos do projeto e de todas as dificuldades impostas para se conseguir uma solução de rastreamento robusta e comum para a diversidade de cenários outdoor envolvidos. No entanto, essa não idealidade entra como um ponto positivo no propósito de testar o detector de anormalidades construído aqui, pois complementa sua avaliação de desempenho mesmo na condição de uso de anotações inexatas. Informações geradas por meio de *datasets* como esses tornam-se úteis na medida em que podem simular os resultados daquela parte de rastreadores não robustos e ainda alvos de pesquisa nessa área. Sendo assim, os *datasets* do LOST trouxeram outras variabilidades importantes para consolidar a efetividade do uso do modelo proposto aqui para aquelas abordagens apoiadas em cenários do mundo real.

As sequências e respectivas anotações de vídeo do LOST são únicas, ou seja, sem a diferenciação de conjuntos destinados para as fases de treinamento ou testes. Isso coloca um outro atrativo para adoção desse *dataset* pois, assim como no Ped2, todas as movimentações dos objetos acontecem naturalmente sem o uso de encenação. Então, para efetivar o uso do LOST, as sequências de vídeo escolhidas são manipuladas através da filtragem e anotações complementares do rastreamento de modo a manter somente os trajetos mais coerentes e longos. Vale destacar que em todas as sequências de vídeo encontradas nas bases de dados das 25 cameras do LOST, somente algumas são adequadas para avaliar modelos de detecção de anormalidades visto que situações anormais nesses vídeos são raras. Assim, durante a investigação, optou-se pela adoção dos vídeos numerados como 1, 14 e 17 onde são encontradas algumas situações tais como o surgimento de objetos não usuais (bicicletas, motos ou animais pequenos) ou ainda movimentos com mudanças de velocidade e direção (pessoas correndo, carros parando, erros de rastreamento como na inversão repentina da associação de objetos com o trajeto, entre outros).

3.1.3 A Anotação de Vídeo

Como mencionado anteriormente e detalhado na subseção 2.3.4 do capítulo 2, o *dataset* Ped2 da UCSD foi anotado com a VATIC de forma completa para produzir um conjunto de amostras de vetores de anotações equivalentes de um rastreamento ideal. Por outro lado, o *dataset* LOST que já disponibiliza anotações de vídeo não robustas, precisou ser submetido a uma seleção das anotações visando eliminar vetores incoerentes ou espúrios decorrentes das imperfeições do algoritmo de rastreamento realizado por Abrams et al. (2012). A não robustez mencionada, considera que a base de informação das anotações contém amostras que refletem os tradicionais problemas enfrentados pelos algoritmos de rastreamento como oclusões parciais e totais, objetos segmentados com mais de um *bounding box*, má associação do trajeto com o objeto, inversão de associação de trajetos durante cruzamento de objetos, surgimento de falsos *bounding box*es devido a falha na segmentação do movimento que surge em variações de iluminação de cenas entre outros.

Para os dois *datasets* adotados, informações de duas variáveis do vetor de anotação necessitaram uma referência comum de classificação: O tipo de objeto (definido como v) e o tipo do seu movimento global. Essas classes foram identificadas por números de acordo com as Tabelas 1 e 2. Números omitidos na sequência da Tabela 1 ficaram reservados para a classificação de outros tipos de objetos não observados nos vídeos adotados para a análise. Essas identificações passam a fazer parte de um vetor de anotações 7-dimensional definido para ser utilizado como referência de entrada no modelo de análise de movimento. Assim, as demais informações que compõe o vetor de anotações são: as coordenadas do centroide do objeto dentro da resolução do *frame*, a largura e altura do *bounding box* dos objetos e o timestamp dos *frames* na sequência. Dessa forma, qualquer vídeo anotado, que disponibilize as 7 dimensões destacadas aqui, pode ser utilizado como base para avaliar o detector de movimentos anormais proposto neste trabalho.

Embora não seja atribuição da DMA identificar qual tipo de anormalidade está em curso, decidiu-se ampliar a segmentação em classes que representam conjuntos diferentes de movimentação. Assim, esses *datasets* também tornam-se apropriados para futuros trabalhos como aqueles destinados para análises mais específicas como a **identificação de comportamento anormal**. Portanto, qualquer tipo de movimento classificados como 1, 2 ou 3 implica que o modelo discutido aqui deve inferir com uma **detecção** de anormalidade mas sem destacar que tipo ela é, embora isso seja possível em futuras implementações uma vez que as anotações em vídeo já estarão disponíveis. As classificações genéricas previstas para os tipos de movimento anormal de objetos da Tabela 2 estão relacionadas com as seguintes situações:

Tabela 1: Classes de tipos de objetos observados.

tipo de objeto (v)	**descrição**
1	1 pessoa
2	1 pessoa + objeto
4	grupo de 2 pessoas
5	grupo de 3 pessoas ou +
6	pessoa usando skate, patinete, patins
7	bicicleta
8	moto
9	animais pequenos
10	automóvel/SUV
13	VAN/caminhões de pequeno porte
14	ônibus
16	caminhões de grande porte

Fonte: Elaborado pelo próprio autor.

Tabela 2: Classes de tipos de movimento de objetos observados.

tipo de movimento de objeto	**descrição**
0 (ou N)	Normal - trajeto usual
1	Anormal - trajeto não usual
2	Anormal - em local não usual
3	Anormal - objeto não usual

Fonte: Elaborado pelo próprio autor.

- Objeto realiza um **trajeto não usual**: quando objetos tomam direções ou sentidos significativamente diferentes dos usuais em qualquer momento durante seu trajeto. Estão inclusos aí movimentos circulares, zigue-zague, movimentos locais e globais atípicos;

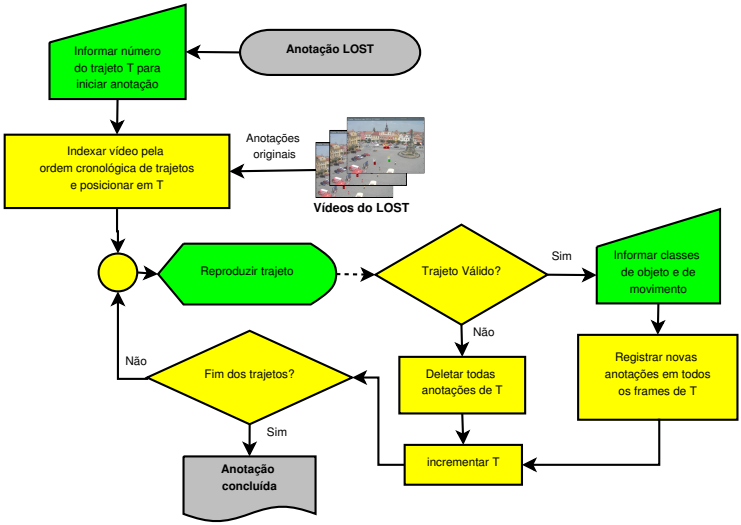
- Objeto se encontra em um trajeto **em local não usual**: pedestres andando na rua ou automóveis ou motos andando em calçadas ou na contramão;

- Movimento de um **objeto não usual**: objetos diferentes dos usuais

como animais, motos, bicicletas, movimentos em velocidades diferentes como pessoas correndo, automóveis parando entre outros.

Especificamente, para realizar a anotação complementar dos vídeos do LOST, foi necessário elaborar uma ferramenta própria para filtrar trajetos ou objetos mal formados e ainda acrescentar as dimensões ausentes na anotação original. A Figura 17 mostra o fluxograma da ferramenta que foi usado como base para a criação do programa desenvolvido na linguagem de script do MATLAB.

Figura 17: Modelagem da etapa de DMA.



Fonte: Elaborado pelo próprio autor.

Embora seja automatizada a varredura das anotações existentes para inserção de novas ou eliminação de outras, o processo de ajuste das anotações de cada trajeto é feito manualmente. Segmentos do vídeo que contém todos os *frames* relacionados com o mesmo trajeto são reproduzidos para melhor conduzir as novas anotações. Apesar da ferramenta ser lenta devido ao processo cíclico na reprodução de cada movimento, ela revelou-se muito útil para o ajuste das anotações de qualquer sequencia de vídeos do LOST permitindo a segmentação de vídeos de treinamento e teste.

3.1.3.1 Análise Local e Global do Movimento

Gryn et al. (2005) e Basharat et al. (2008) definiram que movimentos locais estão ligados aos padrões definidos de vetores locais onde, comparado com um modelo, se conclui sobre seu comportamento. Já o movimento global está ligado aos padrões definidos de distribuição espaço-temporal dos vetores locais. Ou seja, o movimento global é visível em um plano 2D o qual pode ser representado por uma infinidade de caminhos entre pontos ou regiões bem definidos em um modelo de cena, como uma função do tempo.

Na abordagem apresentada aqui, após o treinamento, o detector de movimentos anormais deve ter a capacidade de inferir sobre movimentos atípicos locais ou globais no momento em que eles estão ocorrendo. Os movimentos atípicos locais são aqueles onde mudanças para a próxima localização não são reconhecidas como normais em velocidade ou posição daquele tipo de objeto. Os movimentos atípicos globais já levam em consideração a análise de um segmento completo de deslocamento onde se observam distorções significativas das previsões aprendidas para cada tipo de objeto. Essa é uma contribuição importante que também foi explorada pelo trabalho de Basharat et al. (2008) pois uma avaliação local do movimento pode ser considerada normal a cada próxima observação mas pode se mostrar atípica em uma sequência mais longa de observações do mesmo objeto.

3.1.4 Movimentos Anormais

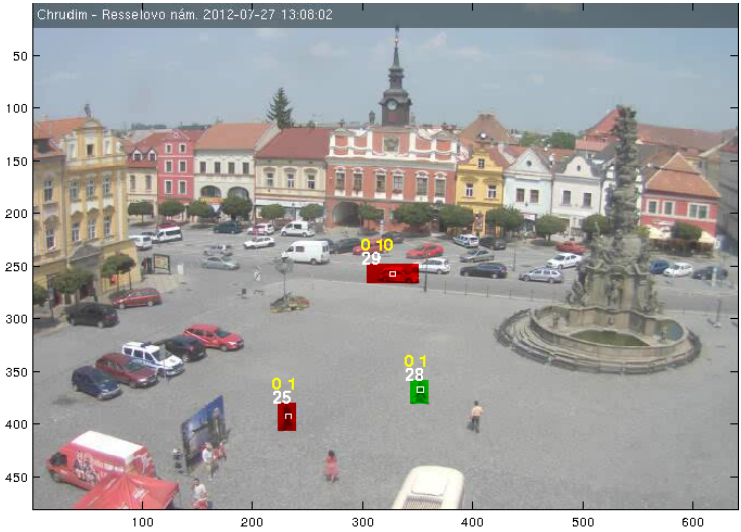
Em essência, movimentos anormais são eventos aleatórios e raros e desse modo, as sequências de vídeo captadas de cenas reais para se transformarem em *datasets*, precisam ser de alguma forma editadas para separar trechos específicos onde esses eventos ocorrem. No caso do *dataset* Ped2, os vídeos de testes seguiram esse critério.

Já nos vídeos do *dataset* LOST, como as capturas das imagens e as anotações correspondentes ocorrem sempre na mesma meia hora de cada dia, mesmo pré-selecionando e editando os vídeos, é difícil encontrar situações de anormalidades suficientes para construir as sequências de teste. No entanto algumas inconsistências contidas nas anotações, resultantes da simplicidade do algoritmo de rastreamento implementado pelos seus autores, podem ser utilizadas como simulação de anormalidades do movimento de objetos. Exemplos destas anomalias são aquelas que produzem mudanças repentinas no rastreamento, as quais ocorrem a partir do erro de associação entre objetos que entram em oclusão durante o movimento. Neste caso, um trajeto que começa associado a um automóvel, pode terminar associado com uma pessoa

a partir da região que ocorre a oclusão entre eles. Este tipo de anormalidade observada nas anotações originais através da ferramenta representada pela Figura 17, é rotulada com uma classe de movimento anormal do tipo “1”, de acordo com a Tabela 2.

A visualização da rotulação da movimentação dos objetos seguiu um padrão similar àquele feito pela ferramenta VATIC. A Figura 18 mostra esse detalhe usando uma amostra de um *frame* reproduzido durante a fase de teste do vídeo 1 do LOST. Na figura, 3 objetos estão rotulados com números sobre os *bounding box* correspondentes, de acordo com a Tabela 2 (acima e à esquerda) e a tabela 1 (acima e à direita). O número impresso em branco representa o número do trajeto na base de dados das anotações. Quando há uma DMA, o *bounding box* é preenchido com a côr vermelha e do contrário, em verde. Nesta amostra, observa-se que devido ao contexto, há outros possíveis objetos se movendo mas os mesmos não estão sendo rastreados. Isso se deve ao apagamento de trajetos que foram convenientemente realizados pela ferramenta proprietária de ajustes de anotações elaborada especialmente para o LOST.

Figura 18: Amostra da rotulação de três objetos de um *frame* do vídeo 1 do LOST.



Fonte: Resultado do detector elaborado pelo próprio autor.

3.1.5 Modelo de Aparência

Como pode-se observar, a anotação feita usando a Tabela 1 já caracteriza completamente um objeto por uma representação direta através de um número. Isso é útil no sentido que evita a tarefa de construir um classificador robusto equivalente e ainda simplifica a quantidade de informação para ser tratada nas próximas etapas. Tarefa essa que também ainda é alvo de pesquisa.

Vale lembrar que as anotações realizadas nos *datasets* Ped2 e LOST carregam as informações de largura (w) e altura (h) dos *bounding box* de cada objeto. Essas informações de alguma forma podem caracterizar, mesmo que de forma pobre, um tipo de objeto no cenário, e evitar assim o uso de um classificador. Isso foi feito por exemplo no trabalho de Basharat et al. (2008). Considerando um cenário que predomina veículos e pessoas, essas variáveis podem ajudar a diferenciar uns dos outros pois a largura e altura desses objetos são bem discriminantes. Obviamente que as tarefas dos estágios anteriores responsáveis pela segmentação de objetos, precisam de robustez para manter uma relação real dessas dimensões com os respectivos objetos. Além disso muitas influências podem provocar distorções severas em w e h como sombras dos próprios objetos, falta de contraste com o fundo estático da cena, variações de iluminação, ângulo de captura da imagem, abertura de membros no caso de pessoas, entre outras.

De qualquer forma, o modelo tratado neste trabalho permite avançar para avaliar seu desempenho usando no lugar uma única variável que representa o tipo de objeto, ao invés do par w e h já disponíveis nas anotações. Outros passos podem ser dados usando modelos de aparência mais complexos usando descritores de cor, textura, forma ou *key points* de um SIFT.

3.1.6 Quantidade de Trajetos

O número de trajetos utilizados para a fase de treinamento é essencial. Dependendo do cenário monitorado, poucos trajetos devem produzir amostras dispersas na ROI e assim dificultando a convergência ou até mesmo o uso de GMM. Por outro lado um número excessivo de amostras eleva o custo computacional e o tempo de convergência sem contribuir de forma relevante na formação das *pdf*. O número mínimo de trajetos fica então dependente das concentrações de movimentação no cenário. Assim, cada vídeo selecionado para o treinamento teve um número convenientemente escolhido de modo que o número de amostras fosse equilibradamente distribuído na ROI.

3.1.7 Homografia

A homografia ou transformação de perspectiva através de um algoritmo de calibração da câmera ofereceria uma contribuição importante somente se a análise de trajetos necessitasse levar em consideração as posições precisas dos objetos sobre um plano real (*ground-truth*) da cena projetado sobre o plano 2D de captação da câmera. Para o propósito deste trabalho, é necessário conhecer somente em que região do plano 2D o centroide de um objeto está, sem que seja necessário conhecer sua real coordenada no cenário monitorado.

Vale lembrar que em sistemas legados de videovigilância as imagens são captadas por uma diversidade de tipos de câmeras e cenários que tornam impraticáveis a realização da homografia e por conta da irrelevante contribuição, dispensável.

3.1.8 Calibração de Câmeras

Naturalmente que um sistema legado de videovigilância é formado por câmeras sem padrões de mínima ou máxima resolução ou outros requisitos que influenciam na sua qualidade como os parâmetros intrínsecos e extrínsecos. No entanto espera-se que os *pixels* captados de cada *frame* tenham uma representação plausível ou com pouca distorção de um objeto enquanto este atravessa o campo de visão da câmera. Estas distorções podem ser determinantes no projeto dos classificadores pois dependendo da distância em que se encontra um objeto alvo da câmera, os descritores podem ficar completamente diferentes. Mas esse é um problema que não afeta o modelo proposto.

Existe um certa "folga" para contemporizar não só essas distorções mas também outras causadas pelas variações de luminosidade, as imperfeições de escala devido as imperfeições das lentes. A premissa aqui então admite que as câmeras utilizadas possuem boas lentes e ajustes adequados de foco, zoom e contraste, bem como uma boa fixação e local de instalação que evita ruídos de vibração mecânica.

A resolução é um aspecto somente ligado ao tamanho mínimo do objeto uma vez que qualquer que seja o tamanho do objeto detectado ele sempre será a cada *frame* representado pelo w e h de seu *bounding box*.

3.1.9 Treinamento Offline do Modelo GMM

Para o treinamento das amostras foi adotada a função *emgm.m*[3] que implementa um algoritmo EM baseado na proposta de Bishop (2006). Essa implementação se comportou com melhor precisão e velocidade na busca dos parâmetros das *pdf* quando comparado com outra proposta por Figueiredo e Jain (2002). Ambas propõe alternativas para tornar viável o uso do EM, contornando suas desvantagens quando ele é utilizado na sua forma padrão. Os autores apresentam algoritmos que são capazes de selecionar o número de componentes (de agrupamentos) de forma automática (não supervisionada) e sem a necessidade de cuidados na inicialização dos modelos de misturas finitas de dados multivariados. O contorno destas desvantagens conduz o uso desses algoritmos para as aplicações de videovigilância, as quais não exigem dependência de treinamento ou rotulação prévia das amostras. Estas condições são fundamentais para permitir uso em modo *on-line*, ou seja, que permite manter atualizado os parâmetros da *pdf* através de um retreino disparado pela presença de novas amostras.

É importante destacar que no âmbito do treinamento do modelo de aprendizagem elaborado neste trabalho, ele ocorre de forma supervisionada e sua metodologia será detalhada na seção 3.4.

3.1.10 Custo Computacional

A complexidade computacional é uma forma de comparar o desempenho entre diferentes algoritmos e que também proporciona uma estimativa do custo computacional que está sendo tratado no presente trabalho. Dentre as medidas possíveis de complexidade, o tempo é a métrica mais adequada para avaliar o desempenho de soluções algoritmicas voltadas para análise de vídeo, especialmente porque uma complexidade de tempo maior pode inviabilizar propostas voltadas para aplicações em tempo real. Nas abordagens baseadas em movimento, onde o processamento das instâncias é contínuo, a complexidade de tempo tem importância fundamental. Como exemplo, na proposta dos autores Haque e Murshed (2012), eles destacam que a complexidade computacional associada não é maior do que a de um processo de extração de fundo de cena baseado em *pixel*. Já Shi et al. (2010) conclui que a complexidade é $\Theta(N^2 logN)$ para todos seus processos onde N é o equivalente ao fator de grade discutido aqui.

Nas abordagens baseadas em rastreamento com treinamento off-line

[3] disponível em *http://www.mathworks.com/matlabcentral*, acesso em 04/08/2014.

como é o caso deste trabalho, uma vez que o modelo é treinado, os resultados são calculados em $O(M)$ onde M depende do número de objetos no *frame*, daqui adiante definido como τ . Isso significa que o número de execuções de operações básicas é fixo e portanto o tempo total é limitado por uma constante que tem valor dependente do número de transições e do número de objetos em movimento em cada *frame*.

3.1.11 Redução de Dimensionalidade

Diante da meta de simplificar a análise de movimento, foi adotado para o treinamento do modelo, apenas 3 dimensões nos vetores de transição definidos a partir dos vetores de anotação disponíveis nos datesets de referência, são eles: o número da região onde o centroide está, o tipo de objeto e seu *timestamp* no quadro. A redução das dimensões desvia a solução apresentada aqui do problema conhecido como a "maldição da dimensionalidade" ou *curse of dimensionality* discutido por Bishop (2006). Obviamente que esta possibilidade se dá devido a premissa de que a instanciação de entrada pode ser reduzida, propositalmente dessa forma.

3.2 MODELAGEM DA CENA

Nesta seção e nas seguintes são detalhados os modelos apresentados no início deste capítulo e arranjados de acordo com o *framework* da Figura 16. O primeiro passo para determinar a construção e implementação do detector de movimentos foi a modelagem da cena onde duas estratégias foram escolhidas para avaliar o desempenho da abordagem: a Divisão da ROI em agrupamentos uniformes e fixos (grade fixa) e uniformes e móveis (grade móvel).

A ideia por trás dessas estratégias é avaliar como o modelo se comporta quando uma grade de regiões se posiciona de modo espacialmente diferente sobre a ROI e também observar os efeitos de desempenho de um modelo quando o fator de grade p_u é variado. A principal motivação de estudar este comportamento com uma análise empírico-científica, foi a ruptura do senso comum de uso de grade fixa na grande parte dos trabalhos de análise de vídeo baseada em região.

Para ambas estratégias, uma vez definido o tipo de grade e o tamanho do agrupamento (ou fator de grade), o modelo de movimento e de aprendizagem seguem desempenhando suas funções.

3.2.1 Divisão da ROI em Agrupamentos Uniformes e Fixos de Regiões

Dentro da revisão bibliográfica realizada, os trabalhos que adotam grades fixas possuem tamanhos de agrupamentos que são sub-múltiplos da resolução do *frame* com o propósito de obter um número inteiro de regiões de mesmo tamanho. Isso pode ser justificado pela conveniência da padronização na modelagem e construção dos algoritmos correspondentes aos métodos e técnicas propostos. Sobre esta estratégia, surgem duas questões importantes: como seria o desempenho da abordagem se o tamanho da região fosse diferente? e qual o melhor tamanho?

Autores como Kwon et al. (2013) se indagaram sobre essas questões e fizeram uma avaliação prévia do modelo com método apropriado para o caso deles. Outros, conforme mencionados no capítulo anterior, de forma heurística, definiram um tamanho que faz emergir a questão sobre a idealidade não só da sua dimensão, quanto do seu arranjo.

Essas questões também serviram como âncoras para realizar toda a modelagem desenvolvida no presente trabalho.

Para o modelo de cena com grade fixa, a ROI é considerada **toda** a área do *frame*. Então o número de regiões possíveis no *frame* vai depender do valor do fator de grade p_u e pode ser determinado com a equação 3.1.

$$g = \left\lceil \frac{R}{p_u} \right\rceil . \left\lceil \frac{C}{p_u} \right\rceil \tag{3.1}$$

A expressão define quantas regiões g um *frame* com resolução $R \times C$ *pixels* será virtualmente dividido. A resolução do *frame* $R \times C$ equivale a resolução de $Linhas \times Colunas$ da câmera que o captou.

3.2.1.1 A Formação da Grade Fixa de Regiões

As regiões $\{r_p\}_{p=1}^{g}$, são numeradas sequencialmente a partir do canto superior esquerdo para o canto inferior direito e de cima para baixo. Isso torna a grade representada por um vetor unidimensional de regiões. Dessa maneira, as duas dimensões que representam a posição 2D do centroide de um objeto, são reduzidas para um escalar o qual representa a respectiva posição no vetor de regiões $\{r_p\}$. Nele serão acumuladas todas as amostras coletadas durante a fase de treinamento ou todos os dados associados diretamente do vetor de anotações de *datasets*.

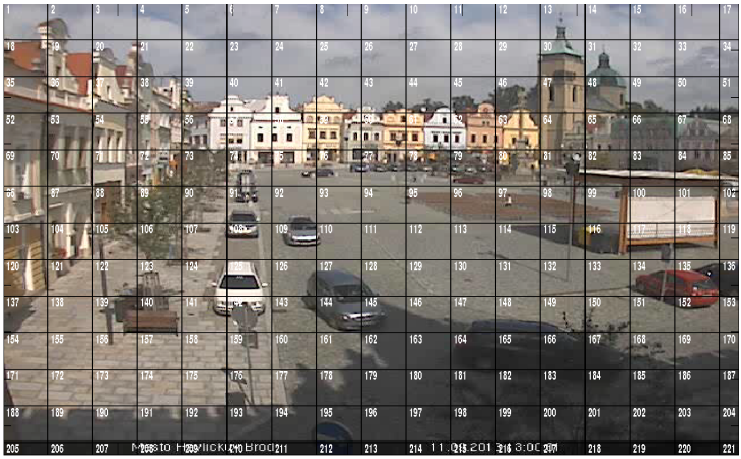
Genericamente, o número da região p onde um objeto se encontra pode ser determinado pela equação 3.2, tomando como base a posição 2D do

centroide (x_u,y_u) onde $\{x_u\}_1^R$ e $\{x_u\}_1^C$.

$$p = \lfloor (x_u - 1)/p_u \rfloor . \lceil C/p_u \rceil + \lceil y_u/p_u \rceil \tag{3.2}$$

A Figura 19 ilustra um exemplo de formação da grade quando é utilizado um fator de grade $p_u = 39$ e $R \times C = 480 \times 640$ *pixels*. A área do *frame* amostrado do vídeo 17 do LOST será transformada para um vetor unidimensional $\{r_p\}_{p=1}^g$ com $g = 221$ elementos.

Figura 19: Rotulação dos números das regiões p na grade fixa sobre um *frame* genérico do video 17 do LOST.



Fonte: Elaborado pelo próprio autor.

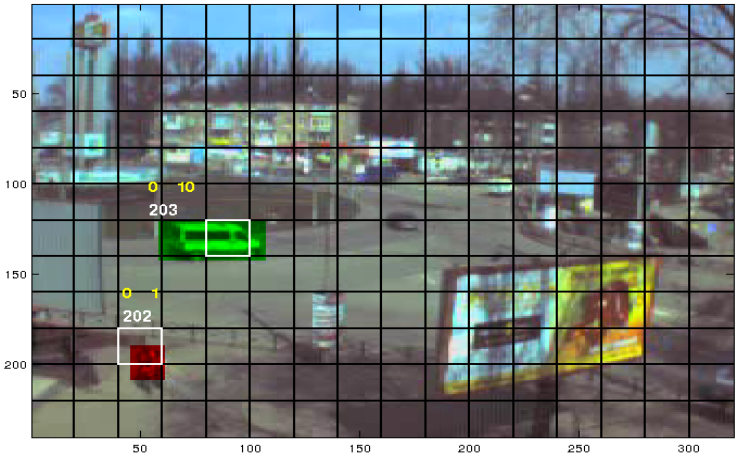
Propositadamente escolheu-se a ilustração da Figura 19 para evidenciar como ocorre a rotulação de todas as regiões no extremo direito e no extremo inferior do *frame* quando o fator de p_u não é sub-múltiplo da resolução do mesmo. Embora isso pudesse ser encarado como um problema para implementar abordagens baseadas em movimento, ele poderia ser facilmente contornado, deslocando a grade para cima e para o lado de modo que todas as regiões mais internas e de mesmo tamanho ficassem centralizadas no *frame*. Somente nesses, ignorando a primeira e última linha e a primeira e última coluna de regiões se aplicaria a abordagem em questão. No entanto, para abordagens baseadas em rastreamento, essas regiões menores só representam regiões menos prováveis de se encontrar ou cruzar centroides de objetos, sendo que, se aí existirem dados suficientes para o tratamento, ele será feito.

Observa-se que a equação 3.2 já prevê a situação quando a resolução

do *frame* não é múltipla da dimensão de p_u. Todas as últimas regiões do lado direito e inferior do *frame* terão agrupamentos menores que $p_u \times p_u$ *pixels*, mas continuaram sendo válidas e numeradas na grade.

A Figura 20 mostra alguns detalhes sobre o uso da equação 3.2 pelo modelo de movimento durante uma fase de teste onde opcionalmente foram adicionados à imagem de um *frame* amostrado do vídeo 14 do LOST, a imagem da grade com $p_u = 20$. Na grade observam-se regiões contornadas em branco indicando que é nela que os objetos tipo 1 e tipo 10 estão associados naquele momento. Então, apesar das áreas dos *bounding box* dos objetos que desenvolvem os trajetos identificados como 202 e 203, estarem sobrepostas em regiões vizinhas, somente as regiões destacadas vão participar da análise por conta do centroide de cada objeto estar dentro dos limites de *pixels* das mesmas.

Figura 20: Sobreposição de áreas do *bounding box* em regiões vizinhas na grade fixa de um *frame* genérico do video 14 do LOST.



Fonte: Elaborado pelo próprio autor.

Somado as imperfeições inerentes ao modelo proposto aqui, algumas inferências incorretas podem ser observadas. Avaliando um pouco mais o trânsito dos objetos sobre a grade fixa, consegue-se identificar situações onde isso acontece. A Figura 21 ilustra exemplos dos problemas intrínsecos neste tipo de modelagem. Dependendo da forma de deslocamento de um objeto sobre a ROI, um fenômeno de vai e vem ou oscilação entre regiões adjacentes pode ser observado quando um trajeto ocorre nos limites de uma ou

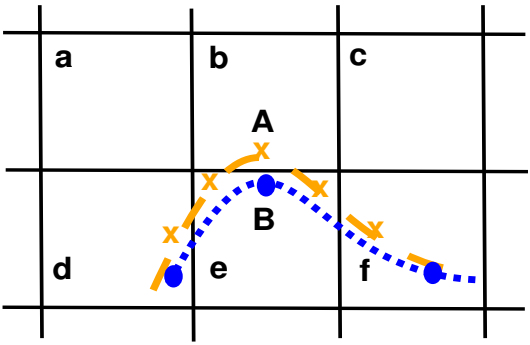
mais regiões. A Figura 21(a) mostra esse efeito que ocorre com o trajeto **A** durante o treinamento. Se a ocorrência é periódica, as transições adicionadas pelo zigue-zague entre as regiões $d \rightarrow e \rightarrow b \rightarrow e \rightarrow f$ são aprendidas como parte do trajeto **A**. Durante a fase de teste, pode existir um trajeto **B**, sutilmente diferente com transições $d \rightarrow e \rightarrow f$ o qual, ao ser avaliado, será detectado como um movimento anormal por não possuir o efeito zigue-zague no meio do caminho. Esta falsa inferência pode acontecer tanto durante o treinamento quanto na fase de teste. Se por exemplo o fenômeno ocorre periodicamente na fase de treinamento repetindo o mesmo padrão, com uma probabilidade maior, um trajeto sutilmente diferente, mas que descreve outro padrão de movimento entre regiões, poderá ser identificado indesejavelmente como um movimento anormal. O mesmo falso-positivo também pode ocorrer quando não se observa esse zigue-zague no treinamento e depois, na fase de teste, mesmo associado a um movimento ligeiramente diferente o efeito é reproduzido.

Outra situação é quando trajetos muito parecidos, mas que possuem amostras distintas capturadas nos limites entre regiões, produzem falso-positivos indesejados. A Figura 21(b) ilustra esse problema que ocorre entre dois trajetos similares mas que possuem padrões de vetores de transição diferenciados durante a amostragem na grade de regiões. As diferenças entre eles são destacadas com cores diferentes onde em verde são regiões comuns, em amarelo região somente associada do trajeto **A** e vermelho região associada somente ao trajeto de **B**. Se somente um dos tipos de trajetos ocorre com mais frequência, esse será aprendido pelo modelo, enquanto que o outro, se avaliado durante a fase de teste, será detectado como um movimento anormal, mesmo sendo muito similar ao trajeto já aprendido.

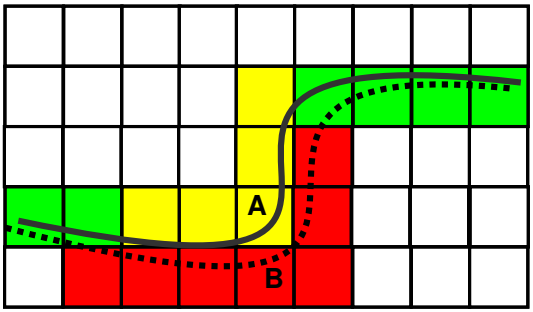
Para contornar esses prolemas, o modelo de movimento pode considerar um período adicional onde o mesmo fica insensível a estas transições, ou seja, realizando uma **contenção de amostragem** durante dois ou mais *frames*, iniciada sempre quando o centroide de um objeto atinge uma nova região. A Figura 22 mostra o resultado de uma contenção de dois *frames*. Neste exemplo, os dois trajetos, mesmo com diferenças em movimentos locais, são considerados totalmente iguais. O modelo agora admite que neste período de contenção, o objeto em questão ainda permanece na última região onde ele foi associado. Como ponto negativo desse paliativo, as amostras de outras transições resultantes de trajetos mais bem ajustados sobre a grade serão perdidas e com elas, a possível perda de representação de outros tipos de trajetos bem como a redução significativa do total de amostras na ROI. Para compensar essas perdas, será necessário incluir uma quantidade maior de trajetos para a fase de treinamento.

No contexto sobre análise baseada em região do presente trabalho,

Figura 21: Situações de falso positivo em trajetos similares.



(a) com transições adicionadas pelo zigue-zague.

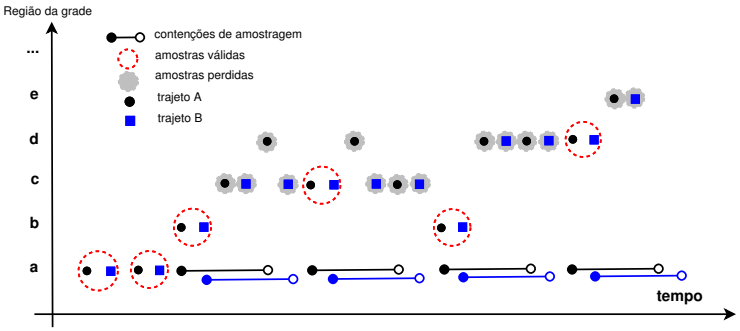


(b) com vetores de transição muito diferentes.

Fonte: Elaborado pelo próprio autor.

similaridades como as descritas anteriormente não deveriam ser encaradas como anormalidades. No entanto são inevitáveis neste tipo de modelagem de cena. A adoção de grades fixas para segmentar a ROI é uma solução simples e portanto atrativa para implementa-la. No entanto a forma geométrica da região não favorece a análise dos dados uma vez que as bordas retilíneas podem desconectar dados vizinhos que eventualmente complementariam e tornariam mais precisa a análise em cada região. Essa perspectiva levou o presente trabalho a extrair do modelo proposto, o melhor desempenho de uma análise sobre as regiões da grade fixa, usando como metodologia, a observação da menor quantidade de falsos positivos quando se varia o tamanho da região. Dessa forma é possível encontrar uma grade com uma área de região que melhor agrupa os dados correlacionados.

Figura 22: Efeito da contenção de amostragem de dois trajetos similares.



Fonte: Elaborado pelo próprio autor.

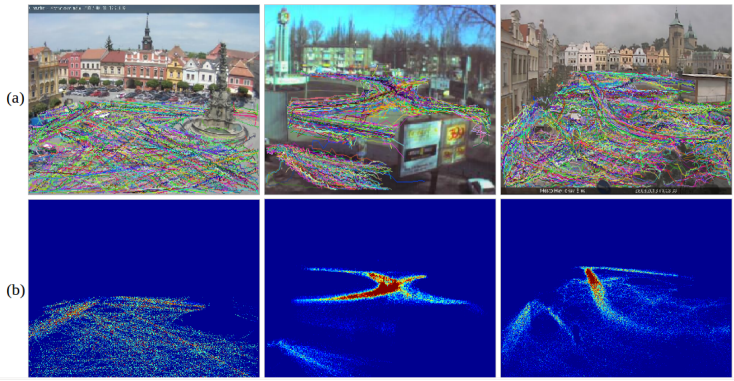
Vislumbrando uma melhora no desempenho do modelo, o próximo passo é não somente encontrar um tamanho adequado de região, como também posicioná-las individualmente em áreas que contenham possíveis agrupamentos de dados com estreita relação. Este modelo de cena também foi implementado e é descrito a seguir.

3.2.2 Divisão da ROI em Agrupamentos Uniformes e Móveis de Regiões

Como mencionado, as transições de objetos entre as regiões em uma grade fixa resulta em erros nas inferências de detecção de anormalidades em trajetórias muito similares. Minimizar esses efeitos com a inclusão de uma contenção de amostragem das transições, leva a outros problemas. Com o objetivo de reduzir estes problemas efeito, procurou-se por outra estratégia que permitisse que as regiões ficassem espacialmente mais encaixadas, envolvendo as amostras dos vetores de transição mais correlacionados e próximos. Para tanto, uma solução implementada foi primeiramente identificar, baseado nas coordenadas 2D disponíveis nos vetores de anotação, a quantidade de amostras desses vetores na área do *frame*. Assim, encontram-se as áreas mais povoadas de amostras as quais serão consideradas a ROI nesta estratégia. A Figura 23 mostra um exemplo da distribuição de amostras de todos os trajetos dos vídeos 1, 14 e 17 do LOST. Na parte de cima, Figura 23(a), estão representados todos os trajetos simultaneamente impressos em cores diferentes. Na parte de baixo, Figura 23(b), estão representados as quantidades de vetores de anotação por região, considerando inicialmente que cada região é de 1 *pixel* de tamanho. As cores mais intensas, em vermelho, denotam as

regiões com as maiores quantidades.

Figura 23: Ilustração da quantidade de vetores de anotação relativos a todos os trajetos impressos em um *frame* genérico dos vídeos 1, 14 e 17 do LOST.



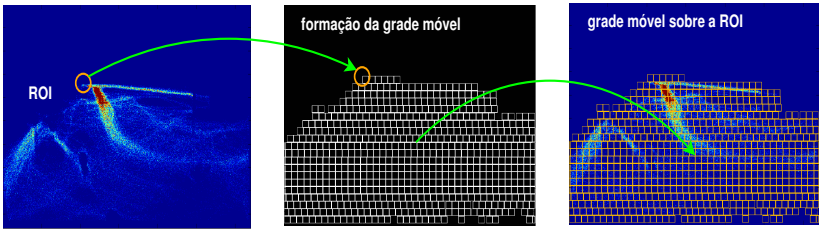
Fonte: Elaborado pelo próprio autor.

Cada amostra do vetor de anotação faz parte de um conjunto de transições de cada trajeto que irá gerar muitas amostras de vetores de transição conforme o modelo de movimento que será detalhado na próxima seção. Então, o número de amostras de vetores de anotação passa a ser uma referência para definir uma região que agrupará vetores de transição fortemente relacionados. Com o objetivo de manter requisitos comuns para estabelecer critérios de comparação, essas regiões também terão uma área quadrada com lado medindo p_u *pixels*.

A grade móvel é construída sempre a partir da primeira região que possui um definido critério mínimo de amostras de vetores de anotação. Essa primeira região terá seu canto inferior direito posicionado sobre o primeiro *pixel* do *frame* o qual levou a atingir ou ultrapassar 5 amostras. Esse valor foi definido como critério mínimo e utilizado para todos os experimentos pelo fato dele produzir quantidades suficientes de vetores de transição que garantem a convergência do GMM na maioria das vezes. Assim, a primeira região da grade é rotulada como "1" e será a referência para o posicionamento da região de número "2" logo à direita assim que essa encontrar a quantidade mínima de amostras definida anteriormente. O processo se repete sempre à direita até encontrar o limite do *frame* ou não encontrar mais amostras de anotação que atendam o critério mínimo. A numeração ocorre sequencialmente da mesma forma como foi feito na estratégia com grade fixa. O processo se repete para uma próxima linha da grade a partir da esquerda do *frame* e logo abaixo da

sequência de regiões anteriores. O processo de montagem da grade móvel continua o mesmo realizado anteriormente até que se conclua todas a área do *frame*. O resultado disso é uma grade com regiões uniformes de lado p_u com formato que se confunde com a ROI onde existem efetivamente amostras para seguir com a análise. Portanto cada região da grade fica móvel para se encaixar somente onde existem amostras significativas para o modelo. Por esse motivo a palavra "móvel" foi atribuído a esta estratégia que terá grades com posicionamentos diferentes de região para cada valor do fator de grade e de cenário. A Figura 25 mostra as fases desse processo considerando o cenário do vídeo 17 do LOST com $p_u = 10$.

Figura 24: Exemplo do posicionamento da grade móvel sobre a ROI do cenário do video 17 do LOST quando $p_u = 10$.



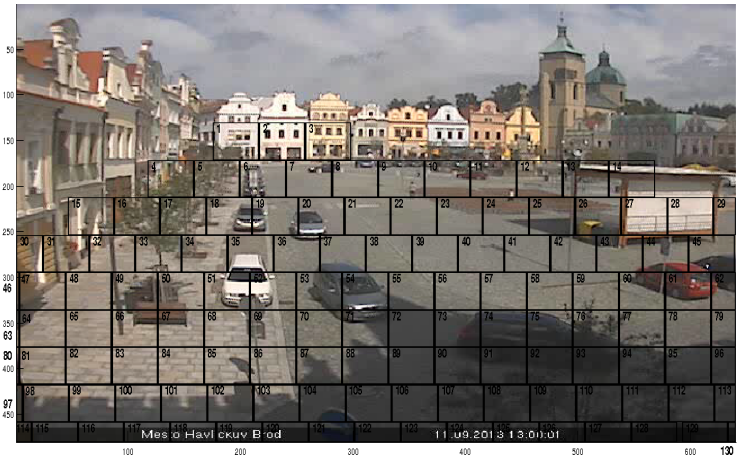
Fonte: Elaborado pelo próprio autor.

Em síntese, a construção da grade móvel tem um ponto de partida na região mais acima e à esquerda da área do *frame* que obedece o critério de número mínimo de amostras de vetores de anotação seguindo à direita e para baixo até atingir o último agrupamento possível no canto inferior direito do *frame*.

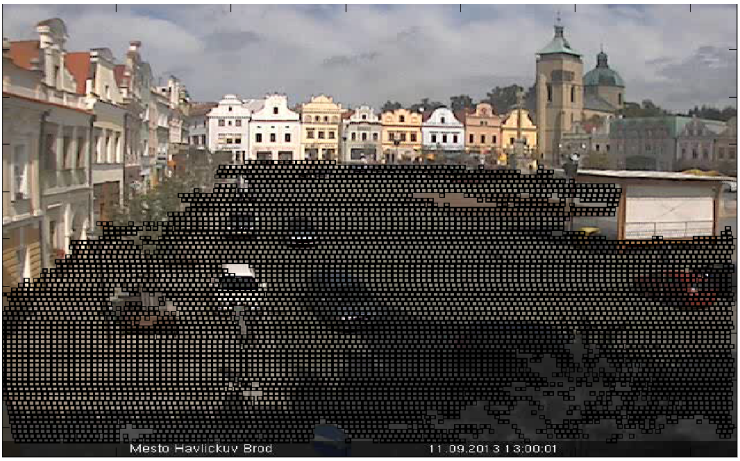
O resultado disso é uma grade com um número menor de áreas e melhor posicionada sobre a ROI, como mostrado na Figura 25(a) a qual já detalha a numeração de regiões na grade e na Figura 25(b) que apresenta um fator de grade bem menor. Nessa última fica claro o posicionamento das regiões sobre as áreas da ROI onde existem número de amostras suficientes de acordo com a distribuição mostrada na Figura 23(b) do vídeo correspondente. Cada região numerada, tem seu posicionamento 2D no *frame* guardado na base de dados do modelo de cena para que a partir dele, se possa determinar em que região o centroide de cada objeto está posicionado em cada transição. Essa é a única forma de determinar a posição da região já que não seria trivial encontrar uma expressão para definir a numeração genérica das regiões p, como foi feito na grade fixa.

Os modelos de movimento e de aprendizagem para a rede móvel são

Figura 25: Exemplos de uma grade móvel sobre um *frame* genérico do video 17 do LOST.



(a) Numeração das regiões quando $p_u = 41$.



(b) Posicionamento das regiões quando $p_u = 5$.

Fonte: Elaborado pelo próprio autor.

os mesmos usados para a grade fixa. A quantidade de regiões é menor do que o modelo com grade fixa, se comparado com o mesmo valor de p_u. A falta de uma expressão genérica para se definir a região acaba exigindo mais algumas

rotinas de cálculo que levam a estratégia a possuir um custo computacional maior, especialmente na fase de treinamento.

3.2.3 Considerações Sobre a Análise Baseada em Região na Abordagem Baseada em Rastreamento Proposta

O vetor de anotações de entrada para o detector de movimentos anormais fica reduzido a um vetor de transições que representa um modelo espaço-temporal com três dimensões (p, v, t) onde t é o valor temporal timestamp no rastreamento do objeto, v é o tipo do objeto anotado e p é a posição na grade de regiões. A distribuição dos dados desses vetores no plano 3-dimensional tende a ficar menos esparsa devido a estreita relação desses dados com todos os movimentos que atravessam a mesma região. A expectativa de tornar as regiões mais povoadas de amostras misturadas, fortemente relacionadas e portanto menos esparsas, melhora significativamente a precisão e convergência de modelos estatísticos como o GMM (FIGUEIREDO; JAIN, 2002).

A redução de esforço computacional mediante redução da dimensionalidade de vetores e de amostras de treinamento implica tornar viáveis as aplicações onde os processos possam ocorrer em tempo real. O uso de grades fixas ou móveis também viabiliza o uso dos modelos avaliados no presente trabalho em sequências de vídeo com resoluções maiores, ainda raras de encontrar na revisão bibliográfica feita até aqui.

Esta é uma estratégia particular deste trabalho que visa estabelecer uma quantidade menor de possibilidades de estados para localização dos objetos, agregando mais amostras inter-relacionadas por região p. Essa redução de estados também traz benefícios como a estabilidade de localização dos objetos que evita o tratamento computacional de transições com pouca ou nenhuma contribuição na avaliação do movimento global.

Outro detalhe importante é que as movimentações do centroide do mesmo objeto dentro das regiões não produzem amostras de vetores de transição e portanto, quanto maior a área da região p menor a quantidade de amostras que serão geradas para serem usadas na fase de treinamento. Relação essa buscada como um dos alicerces deste trabalho.

3.3 MODELAGEM DO MOVIMENTO

Como já mencionado na seção 3.1.1, foi tomado como ponto de partida do presente trabalho, o modelo de movimento encontrado em Basharat et al. (2008) aplicando porém, os modelos de cena discutidos na seção anterior.

A ideia por trás do modelo proposto pelos autores seria mais oportuna se não fosse pela estratégia de engessar a análise em cada *pixel* como região.

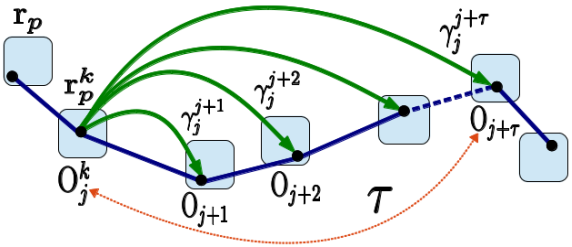
Como ilustrado na Figura 16 no início deste capítulo, o modelo de movimento proposto participa tanto na fase de treinamento quanto na fase de testes do detector de anormalidades. Na fase de treinamento, as coordenadas 2D do centroide são usadas como referência para armazenar em cada região por onde ele passa, informações que irão representar o histórico de uma janela de transições do seu trajeto. Notadamente que essa mesma região poderá ser parte do caminho de outros trajetos acumulando assim, uma quantidade cada vez maior de dados.

A janela de transições, definida aqui como τ, representa o quão longo deve ser observado e avaliado cada trajeto. A premissa para existir uma transição é considerar que um objeto transitou para uma região diferente na observação seguinte. Cada trajeto é sempre analisado dentro do limite definido como τ, independente da quantidade de transições que ele tenha. Quanto mais alto o valor de τ, mais específica será a análise no sentido de inferir sobre movimentos mais complexos.

O *dataset* anotado, oferece um conjunto de n trajetos T para cada vídeo representados como $\{T_i^k\}_{i=1}^n$ e $k \in \mathbb{N}^*$ é o conjunto de *frames* k onde cada objeto é amostrado. Cada *frame* k tem um bem definido *timestamp* t no vídeo e $t \in \mathbb{R}^+$. Então T_i^k representa um conjunto de m observações do mesmo objeto, $T_i^k = \{O_j^k\}_{j=1}^m$. Cada observação é um conjunto de vetores de transição $O_j^k = \{\gamma_j^{j+a}\}_{a=1}^{\tau}$, onde $\gamma_j^{j+a} = (p,v,t)^T$ é o conjunto de vetores de transição das próximas τ transições. Eles contém o registro temporal contínuo t (*timestamp*) do tipo de objeto v, na região p da grade. A Figura 26 mostra as futuras transições observadas a partir de qualquer objeto no *frame* k. Elas produzem amostras adicionais na região onde o objeto está atravessando. Para qualquer trajeto observado em qualquer *frame* k, uma janela de amostragem até τ é realizada. Todos os vetores de transição até $\gamma_j^{j+\tau}$ são associados como amostras na região do ponto de observação O_j^k.

A redução da dimensionalidade do vetor γ_j para três dimensões foi propositalmente adotada para também contribuir na redução do esforço computacional, apesar de que essa consequência não ficará evidenciada de modo cristalino nos resultados do modelo, os quais serão descritos no próximo capítulo. A ênfase dada sobre custo computacional enfatiza o número de amostras de vetores de transição como métrica base. No entanto, vale lembrar que a característica de recursividade, comumente utilizada nos algoritmos que tratam dados multivariados, são focos importantes no consumo de recursos computacionais. Então, reduzir dimensionalidades do vetor de transições também trás contribuições no quesito eficiência de um modelo.

Figura 26: Detalhe da construção do modelo de movimento.



Fonte: Modelo proposto por Basharat et al. (2008) e adaptado pelo autor.

3.4 MODELAGEM DA APRENDIZAGEM

Diferente dos modelos estudados na revisão bibliográfica feita até aqui, o modelo de aprendizagem possui duas fases de treinamento com objetivos bem definidos: a primeira fase compreende a **geração de amostras** de vetores de transição para cada região a partir das sequencias de vídeo de treinamento, realizadas com o apoio do modelo de movimento discutido na seção anterior. Assim que as amostras estiverem disponíveis, os parâmetros das *pdf* por região serão levantados pelo modelo GMM treinado por um algoritmo EM. Na segunda fase, agora com os vídeos de teste, ocorre o **levantamento do melhor limiar** de probabilidade que consegue separar trajetos normais de anormais através do uso de curva ROC. Outros detalhes dessas fases serão abordadas mais adiante.

Todas as *pdf* devidamente parametrizadas, são armazenadas nas bases de dados do detector de movimentos anormais proposto aqui. Para que se garanta o início das fases de treinamento é necessário uma quantidade suficiente de trajetos que revelem uma ROI povoada com amostras suficientes de vetores de transição. Para se alcançar esse objetivo há dois caminhos: **i)** manter longos períodos de observação em cenários que possuem pouca movimentação de objetos e/ou possuem distribuição esparsa na movimentação dos mesmos. Esses são os casos dos vídeos 1 e 17 do LOST; ou **ii)** realizar curtos períodos de observação em cenários que possuem uma densidade relativamente grande de objetos se movimentando. Estes são os casos dos vídeos 14 do LOST e dos dois *datasets* da UCSD. Em qualquer situação os trajetos devem ser válidos e longos o bastante para que nos casos de grades com fator de grade maior, os trajetos podem deixar de existir pois eles podem iniciar e terminar na mesma região, ou seja, não apresentando transições.

As tabelas 3 e 4 resumem as informações dos *datasets* utilizados no treinamento e na avaliação do detector proposto aqui. Foram utilizados tipos diferentes de vídeos não só em cenários ou resoluções, mas também em tempo de vídeo e quantidade de trajetos disponíveis.

Tabela 3: Informações sobre os vídeos de treinamento dos *datasets*.

Dataset TREINAMENTO	**LOST video 1**	**LOST video 14**	**LOST video 17**	**UCSD Ped2**
resolução	480x640	240x320	480x640	240x360
FPS (média)	0,5	8,9	6,1	30,0
minutos de vídeo	240	240	300	1,5
trajetos normais	1190	1755	2990	265

Fonte: Dados levantados pelo autor.

Tabela 4: Informações sobre os vídeos de teste dos *datasets*.

Dataset TESTE	**LOST video 1**	**LOST video 14**	**LOST video 17**	**UCSD Ped2**
minutos de vídeo	240	240	300	1
trajetos normais	1190	1755	2990	182
trajetos anormais	37	32	116	25

Fonte: Dados levantados pelo autor.

Em relação aos vídeos do LOST, a soma dos trajetos normais e anormais, da tabela 4 encontram-se representados graficamente na Figura 23(a). Com um olhar mais atento na distribuição das amostras de vetores de anotação, representados na Figura 23(b), observa-se que o vídeo 1 além de possuir o menor número de trajetos, as amostras correspondentes estão bastante dispersas por uma área relativamente grande do cenário. Isso também ocorre no vídeo 17, apesar de ele possuir uma região no centro do cenário com uma visível área de concentração de amostras. A dispersão de amostras sugere que muitas áreas podem ter dados insuficientes para o treinamento do GMM. Em contrapartida o vídeo 14 é um representante de superpopulação de amostras na maior parte de sua ROI. A escolha desses vídeos dentre os vários disponíveis no LOST foi proposital, especialmente pela diversidade de contextos e distribuição de seus trajetos na ROI. Eles vão ajudar a avaliar o modelo proposto no que se refere a sua capacidade em analisar movimento em condições diferenciadas.

A essência do modelo de aprendizagem é descobrir para cada região p, uma quantidade de *pdf* com os respectivos parâmetros de média (μ) e ma-

triz de covariâncias (Σ) que melhor representem os objetos e a história dos respectivos trajetos que por ela passaram. Dessa forma, na fase de teste, qualquer desvio do movimento usual resulta em valores de probabilidade muito baixos e anormalidades são identificadas. Estando a dependência do modelo ligada especialmente às características do objeto e seu movimento, o restante do cenário e dos outros diversos tipos de objetos móveis são variáveis que não devem influenciar de modo significativo sobre a identificação de anormalidades de movimento sobre os outros. Exemplos dessa independência de contexto foram discutidos na seção 2.6.

Então, qualquer desvio significativo do movimento de costume em todas τ transições irá produzir diferenças quando a probabilidade for calculada implicando assim na detecção das anormalidades. Um movimento anômalo local dentro de um global no trajeto, rotula o evento como anormal. Considerando-se os agrupamentos de dados por região, com uma coleção de vetores de transição de dimensionalidade igual a 3, a probabilidade de cada vetor de transição é determinada pela equação 3.3, onde η_p representa a quantidade de amostras em cada região p e $a = \{1, 2, ...\tau\}$.

$$P(\gamma^{j-a}|(\Sigma,\mu)_{r_p}) = \frac{1}{\sqrt{(2\pi)^3|\Sigma|}\eta_p} \exp^{-\frac{1}{2}(\Sigma-\mu)^T\Sigma^{-1}(\Sigma-\mu)} \tag{3.3}$$

Uma rodada de treinamento do modelo de aprendizagem pode ser resumida no pseudo-código detalhado no Algoritmo 1. O treinamento ocorre em um processo iterativo, realizado em modo off-line para encontrar os valores de p_u que produzem os melhores índices de acertos nas inferências sobre os vídeos de teste começando com $p_u = 1$. Na primeira fase, nas linhas 4 à 13 do Algoritmo, somente os vídeos de treinamento, livres de trajetos com anormalidades participam no levantamento das *pdf* por região. Portanto, este método refere-se a uma aprendizagem supervisionada, uma vez que os dados de treinamento compõe-se por apenas uma classe de trajetos (os normais) (SODEMANN et al., 2012). Na segunda fase, nas linhas 14 a 18, são utilizadas as sequencias de vídeo de teste que contém movimentos normais e anormais devidamente anotados nos vetores de anotação.

A contabilização do total de amostras durante cada iteração é realizada na linha 20 e somente será utilizado como referência para avaliar, através do gráfico montado na linha 22, o custo computacional associado ao desempenho alcançado a cada novo valor de p_u. Essas avaliações serão discutidas nos próximos capítulos, mas por ora, relembrando o modelo de movimento adotado, cada transição provoca $\tau - 1$ amostras de vetores de transição no trajeto até as últimas τ transições. Isso implica que o total de amostras envolvidas em cada avaliação do valor de p_u será aproximadamente um pouco menor do que

Algorithm 1: One Round Training and Best p_u determination

input : Data Annotations of objects v types in t transitions time of n tracks; *grid type*

output: 3-dimensional *pdf* per region; $Best\{SceneThreshold, p_u\}$

Inicialization: $p_{u_{max}} = 30; \tau = \{20;40\}\ targets_n$;
Run *Scene Model* to determine p regions grid positioning;
for $p_u \leftarrow 1$ **to** $p_{u_{max}}$ **do**
    `// 1st Training Step with training dataset`
    for $p \leftarrow 1$ **to** *all grid regions* g **do** $\{\Sigma, \mu, k\}_{r_p} = \emptyset$;
    forall the *n tracks* **do** in each r_p^j grid localization
        foreach $a \leftarrow 1$ **to** τ **do** transitions
            $r_p^{j+a} \leftarrow \gamma^{j+a} = [r_p^j, v, (t^{j+a} - t^j)]$;
        end
        for $p \leftarrow 1$ **to** g **do**
            Run EM algorithm over Gaussian Mixtures in r_p;
            Save learnt *pdf* parameters Σ_p, μ_p, k_p;
        end
    end
    `// 2nd Training Step with test dataset`
    forall the *n tracks* **do** in each r_p^j grid localization
        out_n = estimate *min* probability in $\gamma^{j-a}|(\Sigma, \mu)_{r_p}$ from τ previous transitions;
        Run ROC curve from $\{out_n, targets_n\}$ vectors;
        $ROCefficiency_n \leftarrow threshold_n$;
    end
    $threshold_{p_u}$=max$\{ROCefficiency_n\}$;
    $TotalSamples_{p_u} \lesssim TotalTransitions(\tau - 1)$;
end
Plot $\{TotalSamples; threshold_{p_u}\}$ by p_u;
$BestSceneThreshold$=max$\{threshold_{p_u}\}$;
$Best p_u \Rightarrow BestSceneThreshold$;

o total de transições existentes entre as regiões da grade, multiplicada pelo valor de $\tau - 1$. Então, a medida que o valor de p_u sobe, o número de regiões da grade diminui e consequentemente o número equivalente de transições entre elas, levando a um número de amostras cada vez menor quanto menor a quantidade de regiões.

O valor de p_u é então incrementado em uma unidade e um novo ciclo com as duas fases se repete, chegando a um outro valor de limiar para o novo valor. Todos os dados e variáveis envolvidos em cada iteração são armazenados para realizar a análise. Para melhor visualização, os resultados são plotados para identificar qual dentre todos os valores de fator de grade obteve melhor resultado, buscado aqui como meta do DMA modelado. Associado a esse valor campeão, está o correspondente valor de limiar que será adotado, a partir dessa rodada *offline*, como um classificador binário, até a necessidade de uma nova rodada.

3.4.1 A Construção da Curva ROC

Uma vez que se está interessado apenas na mais alta taxa de acerto de verdadeiros positivos (TPR - True Positive Rate) e a menor taxa de acerto de falsos positivos (FPR - False Positive Rate), foi adotada como métrica de referência a **ROCefficiency** através equação 3.4.

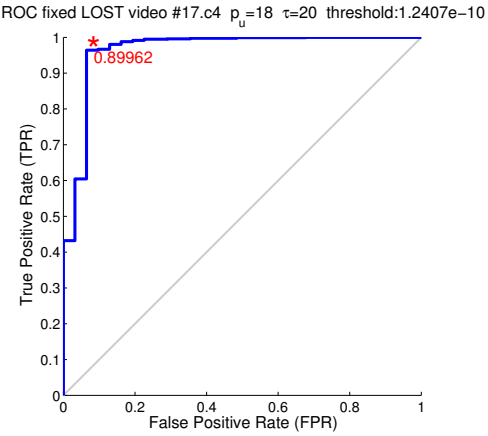
$$ROCefficiency = (TPR - FPR).(1 - \frac{\varepsilon}{total\ tracks})^2 \qquad (3.4)$$

Esta métrica foi particularmente escolhida para estabelecer uma relação de qualidade nas inferências corretas não só sobre trajetos anormais quanto para os normais, considerando um número finito e reduzido de amostras de trajetos. Para classificadores binários, Powers (2011) sugere uma medida de desempenho equilibrada por $(TPR - FPR)$, o qual foi definido por ele como *informedness*. Um número mais próximo do limite 1, indica um melhor índice de acertos para todos os tipos de movimentos observados no cenário, normais e anormais. O valor ε representa o número de trajetos perdidos, que serve como fator de penalidade. Ele procura manter um critério mais justo de comparação entre avaliações que possuem índices próximos de eficiência mas com números de trajetos diferentes. As perdas de trajetos ocorrem por dois motivos: (i) o número de amostras em todas as regiões de transição do objeto não foi suficiente para inicialização do processo de treinamento ou de convergência no treinamento do GMM. O algoritmo utilizado, conforme discutido na seção 3.1.9, dependendo das variabilidades entre as amostras,

requer agrupamentos de pelo menos 30 vetores de transição e (ii) a falta de transições entre as regiões por conta dos trajetos mais curtos que acabam começando e terminando dentro de uma mesma região quando o valor de p_u é incrementado. Nesse ponto observa-se que o valor de τ é importante e por esse motivo, grande parte das avaliações usaram valores diferentes nessa variável, conforme previsto na linha 1 do Algoritmo 1. Sendo assim se as perdas de trajetos forem significativas em relação a quantidade total disponível, elas irão desequilibrar a comparação de desempenho entre resultados com p_u diferentes. Como alternativa de penalização, a equação 3.4 já prevê uma compensação dessa desigualdade.

O melhor limiar encontrado na linha 19 diante da maior eficiência calculada pela equação 3.4, representa o melhor valor de um limiar que separa (classifica) movimentos normais de anormais para cada valor de p_u. A Figura 27 mostra um exemplo da representação gráfica de como o valor ótimo de *informedness* foi encontrado para o vídeo 17 do LOST com valor de $p_u = 18$. Esse ponto é representado pelo asterisco destacado na curva e ao lado dele encontra-se o valor numérico da medida de *informedness* já multiplicada pelo fator de penalização previsto na equação 3.4. Então, o valor indicado pelo asterisco é a métrica *ROCefficiency* para aquele valor de p_u.

Figura 27: Exemplo de uma curva ROC construída durante o treinamento.



Fonte: Elaborado pelo autor.

Para construir essa curva, são necessários dois vetores de mesmo tamanho da quantidade de trajetos em treinamento. O vetor de alvos ou de referência, *target* e o vetor de resposta, *out*. O vetor *target* é um vetor binário

formado a partir do vetor de anotações de vídeo onde um binário "1" é o valor do elemento do vetor que define um trajeto normal e "0" para um trajeto anormal. O vetor *out* compõe-se de valores de mínima probabilidade encontrada com a equação 3.3 dentre todas as transições de cada trajeto. Esses vetores serão as duas referências exigidas para construir a curva ROC prevista nas linhas 16 e 17. Para formar essa curva, todos os valores de probabilidade do vetor *out* são colocados em ordem crescente formando o vetor de limiares ou vetor *threshold*. Cada elemento deste vetor é confrontado com os vetores *out* e *target* para contabilizar quantos trajetos anormais, referenciados em *target* como anormais foram corretamente inferidos até com valores iguais ou menores ao elemento em curso. Dividindo-se essa quantidade de acertos pela quantidade de trajetos marcados como anormais no vetor *target* tem-se a taxa de verdadeiros positivos, TPR e esse valor é uma das coordenadas da curva. De forma equivalente o mesmo procedimento é feito para se determinar a taxa de falsos positivos, FPR, considerando agora quantos erros de trajetos normais, referenciados como normais em *target* são alcançados para limiares acima do valor em curso no *threshold*. No caso, a segunda coordenada da curva usa a taxa de erros em relação a quantidade de trajetos marcados como normais. O processo se repete para todos os valores do vetor de *threshold*. Ao final tem-se uma curva traçada com todos os pontos. Pode-se observar na Figura 27 uma linha referencial diagonal entre as coordenadas (0,0) e (1,1) de TPR e FPR. Ela representa a divisão entre duas regiões onde o classificador se comporta como ineficiente, abaixo da diagonal e eficiente, acima da diagonal e aleatório sobre a diagonal. Pontos sobre a linha diagonal indicam que o classificador não possui qualquer capacidade de inferir sobre a análise e assim, aleatório. Portanto, os melhores valores de limiares de um classificador binário encontram-se no topo superior esquerdo do gráfico próximos ou nas coordenadas (1,0), ou seja, quando o TPR é máximo e o FPR é nulo. O valor 0,89962 mostrado na Figura 27, considerado-se que não há perdas de trajetos, revela que um limiar de 1.2407e-10 consegue identificar corretamente como anormais, aproximadamente 96% do total de trajetos apresentados como tal pois eles possuem limiares de probabilidade abaixo desse limiar. Em contrapartida o mesmo limiar erra aproximadamente 4% dos trajetos normais do total existente.

3.5 MODELO DE TESTE

Uma vez que tanto o melhor valor de p_u e respectivo melhor valor de limiar de decisão são determinados para a cena, qualquer sequencia ou tamanho de vídeo do mesmo cenário que contém vetores de anotações, pode

ser testado. Na fase de teste as sequencias de vídeo para este fim ficam dependentes do *dataset* utilizado. No LOST é possível realizar anotações de vídeos adicionais, diferentes daqueles utilizados nas duas fases do treinamento em vista da disponibilidade oferecida pelos autores do projeto nas correspondentes páginas *web*. Já nos *datasets* da UCSD, a exemplo do que ocorre com a maioria dos *datasets* disponíveis para esse tipo de pesquisa, os testes de desempenho da abordagem proposta aqui se limitam a usar os mesmos vídeos de teste usados na fase 2 do treinamento do modelo de aprendizagem. A forma de visualização das detecções de movimento anormal na fase de testes já foram ilustradas na Figura 18 da subseção 3.1.4. Em uma aplicação real, cada cenário dos vários centralizados em uma monitoração de videovigilância, roda uma instancia do Algoritmo 1 de modo que ele possa ser reexecutado em períodos pré-definidos que englobem perfis diferentes de movimentação do cenário, típicos de cenários reais onde objetos e movimentos correspondentes são distintos ao longo do dia, da noite ou dos dias da semana. Assim, cada round oferece uma grade de regiões com *pdf* parametrizadas através do modelo de aprendizagem apresentado além do limiar classificador de movimentos anormais em como o valor de p_u que maximiza o desempenho da abordagem. Para colocar o modelo em teste utiliza-se o Algoritmo 2 a seguir.

O Algoritmo 2 recebe como entrada as *pdf* parametrizadas pelo Algoritmo 1 além dos valores ótimos de p_u e limiar para a cenário monitorado. Considerando a mesma resolução de *frame*, deve-se utilizar as coordenadas do tipo de grade de regiões utiliza no treinamento omo referência para, sobre ela avaliar, cada transição de cada objeto em cada *frame* de vídeo. A raiz desta fase de teste é o que trata as linhas 7 e 8 onde, para cada transição, um teste é feito para determinar se existe a probabilidade do objeto j estar na região p atual, tendo ele origem e tempo de deslocamento nas τ transições anteriores. Qualquer valor calculado, usando a equação 3.3, menor que o limiar λ conhecido da fase de treinamento, leva a entender que o objeto avaliado não deveria estar naquela região e portanto é encarado como uma transição anormal. A partir desse momento, enquanto o objeto estiver sendo rastreado na cena ele estará sendo identificado por uma cor vermelha em seu *bounding box* ou qualquer outra ação que se deseja implementar visando dar suporte a alarme a pessoa que etá monitorando o vídeo.

Algorithm 2: Test Model after one round Training

input : Tracking with Data Annotations of the objects with v types in τ transitions before; *grid type* with g regions; *pdf* parametrizada de cada região com $\{\Sigma, \mu, k\}_{r_p} and \{p\}_1^g$ from training

output: Abnormal Motion Analys in Video

Inicialization: $p_u = Best\,p_u; \lambda = BestSceneThreshold$;
while $\exists$ *tracking* **do**
  foreach *frame with n transition objects* **do**
    foreach $j \leftarrow 1$ **to** *n objects* **do**
      for $a \leftarrow 1$ **to** τ **do**
        $\gamma^j = (r_p^{j-a}, v, (t^j - t^{j-a}))$;
        if $P(\gamma^j | (\Sigma, \mu)_{r_p}) < \lambda$ **then**
          Put red color in Bounding Box and others actions; **break for**
        else
          Put green color in Bounding Box or no actions
        end
      end
    end
  end
end

Assim, baseado em todo o modelamento previsto e detalhado neste capítulo e em especial o Algoritmo 1, as rotinas foram desenvolvidas e implementadas em linguagem interpretada do MATLAB. Os resultados mais representativos, bem como as avaliações dos pontos positivos ou das distorções encontradas serão apresentados no próximo capítulo.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

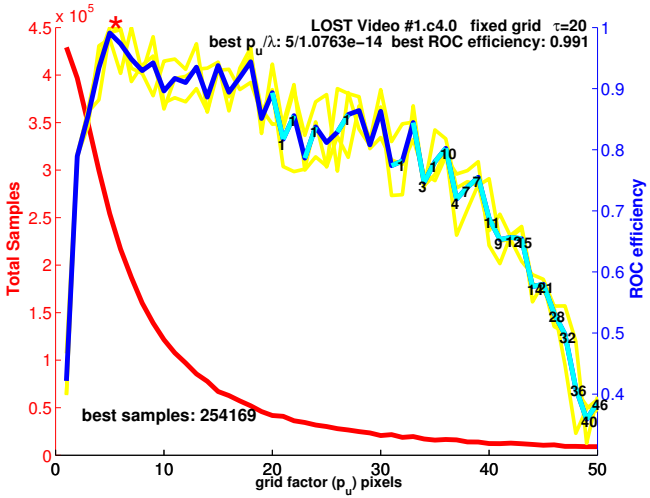
O Algoritmo 1 discutido na seção 3.4 foi a base na implementação e geração de todos os resultados descritos nesse capítulo. Visando sintetizar as principais informações geradas nas simulações das fases de treinamento, foi adotado um modelo de gráficos de curvas conforme exemplo ilustrado na Figura 28 o qual, daqui em diante, é denominado simplesmente como gráficos de desempenho. Os resultados das iterações são demonstrados através de gráficos sobrepostos que objetivam expressar de forma imediata as relações e o comportamento entre o número de amostras de vetores de transição utilizados no treinamento e o desempenho do modelo usando a métrica *ROC efficiency* através da variação do fator de grade p_u.

A ordenada esquerda indica a quantidade de amostras utilizadas no treinamento e o comportamento correspondente, representados pela curva vermelha a qual revela um caimento do total de amostras com o aumento de p_u. Esse comportamento foi comum para todas as simulações pois a medida que se aumenta o valor da área da grade, as transições do objeto que ocorrem dentro dessa área são descartadas. Como consequência, vídeos amostrados com um número maior de FPS tendem a possuir uma queda de amostras exponencialmente maior, pois as próximas transições ocorrem em *pixels* vizinhos muito próximos, especialmente tratando-se de cenários com pedestres. Já a ordenada da direita representa o melhor valor da métrica *ROC efficiency* e o comportamento correspondente, representados pela curva azul a qual apesenta a evolução dos valores encontrados também para cada valor de p_u. Ao longo dessa curva, quando for o caso, são impressos valores que representam a quantidade das perdas de trajetos, apresentadas na seção 3.4.1. Quando as perdas ocorrem, além da indicação colocada ao lado da curva, o segmento correspondente é mostrado em um tom de azul mais claro. O asterisco sobre a curva destaca o valor de p_u que representa o melhor desempenho nas inferências da DMA e o correspondente limiar ótimo para o classificador binário o qual, dentre outras informações, é incorporado ao título superior de cada gráfico.

Por se tratar de um modelo estatístico, é necessário encontrar o valor médio dos comportamentos de *ROC efficiency* em razão de que o treinamento do GMM com EM pode não convergir para cada nova rodada mesmo com o mesmo conjunto de amostras, como também pode convergir para valores de parâmetros da *pdf* com variações significativas em relação a rodadas anteriores. Assim, os valores ótimos de p_u e limiar λ bem como a evolução dos valores de *ROC efficiency* e perdas, são resultados da média de 3 simulações completas de qualquer cenário. O traçado de cada simulação dessas curvas

está ilustrado na Figura 28 na cor amarela. Em alguns casos estas curvas podem não estar destacadas no gráfico, mas ainda assim a curva representará o resultado médio da análise. Por final, no gráfico, logo abaixo das curvas de desempenho, é informado a quantidade de amostras (*best samples*) para o indicado valor de p_u que melhor representou o desempenho da simulação. Durante todas as simulações observou-se que valores de p_u a partir de 30 não produziam mais informações relevantes para a análise na maioria das simulações em todos os vídeos avaliados. Então, tomou-se esse valor quando foi o caso, como limite mínimo para construir a maior parte do gráficos.

Figura 28: O modelo adotado para as curvas resultantes da simulação das duas fases de treinamento do modelo de aprendizagem.



Fonte: Gráficos construídos a partir das aplicações MATLAB implementadas pelo próprio autor.

O modelo de aprendizagem apresentado na seção 3.4 destaca que o treinamento utiliza o próprio vídeo de teste para descobrir o melhor limiar do classificador binário. Nesse aspecto, dentro do que é definido como treinamento supervisionado, duas possibilidades podem ser utilizadas para analisar o desempenho da abordagem: a primeira usa como vídeos de teste, na segunda fase de treinamento, os mesmos vídeos usados na primeira fase porém agregados com trajetos anormais. Essa estratégia é conhecida como análise baseada em reconhecimento e produz resultados que medem o desempenho a partir de uma métrica denominada de **taxa de erro aparente**. A expecta-

tiva de modelos que usam essa estratégia é alcançar taxas de erro em níveis muito próximos de zero, demonstrando assim, a capacidade do modelo distinguir padrões até então não apresentados na fase de aprendizagem. A segunda possibilidade utiliza vídeos de teste com trajetos normais diferentes daqueles que participaram na primeira fase do treinamento e também com trajetos anormais adicionados. Nesse caso a análise produz resultados que medem o desempenho a partir de uma métrica denominada de **taxa de erro verdadeira** e ela mede a capacidade de generalização do modelo (BISHOP, 2006). Para o propósito da abordagem apresentada no presente trabalho, a primeira possibilidade é uma opção que permite realizar uma sintonia mais fina na busca pelo melhor limiar classificador, uma vez que somente os trajetos anormais devem apresentar valores de probabilidades nulas ou muito baixas. Com essa estratégia atenua-se o efeito das incertezas decorrentes do uso do modelo caso ele fosse avaliado pela análise de taxa de erro verdadeira, afetando assim as metas traçadas do presente trabalho. Diante disso, todos os resultados apresentados neste capítulo estarão baseados no modelo de treinamento onde suas duas fases usam os mesmos trajetos normais. Para fins de avaliação mais completa, não será descartada a análise de taxa de erro verdadeira e isso será feito como uma discussão periférica no próximo capítulo.

A usabilidade da aplicação criada para as simulações foi planejada para permitir uma série de combinações de resultados. A meta na implementação foi não só buscar subsídios para constatar a contribuição da abordagem da DMA proposta, como também na ampliação da avaliação do modelo em outras situações e na diagnose de outros comportamentos observados nos dados. Tratando-se de um método empírico-científico, as combinações de simulações para cada vídeo na abordagem foram executadas variando-se valores de p_u, τ, tipo de grade, comprimento de vídeo, tipos diferentes de algoritmos de treinamento do GMM, exclusão de objetos com determinadas dimensões de *bounding box*, tempo de contenção de amostragem e modos de treinamento supervisionado ou não. Entre todas as possibilidades, voltou-se a atenção para aquelas que melhor representaram os objetivos traçados no presente trabalho. Outros resultados não menos importantes, serão discutidos no próximo capítulo.

Nessa linha de raciocínio, as seções seguintes descrevem os principais resultados os quais foram separados por *dataset* e dentro de cada um, foram agrupadas as análises dos tipos de grade utilizado na modelagem da cena, comprimentos diferentes de vídeo e de janela de transições (τ). No início de cada seção, serão detalhadas informações importantes sobre as características dos vídeos e dos números envolvidos dos *datasets* correspondentes, os quais foram utilizados durante as duas fases de treinamento do modelo de aprendizagem. Ao final de cada seção serão resumidos através de tabe-

las, os principais resultados observados nas curvas de desempenho de todas as simulações bem como alguns comentários preliminares sobre as relações entre os comportamentos experimentados. Na última seção deste capítulo, outras impressões e análises discorrem sobre os resultados em âmbito geral da evolução das simulações, frente as expectativas traçadas para a abordagem proposta.

As estratégias de execução das simulações foram adotadas visando a geração de dados comparativos especialmente entre os tipos de grade usados no modelamento de cena. No caso dos vídeos do LOST foram escolhidos aleatoriamente segmentos diferentes de 30 minutos cada. Dessa maneira, com algumas exceções, as simulações foram agrupadas primeiro por tipo de *dataset*, depois pelo período de observação que acumulou as amostras de trajetos e por fim, pela quantidade de transições entre regiões. Em relação ao período de observação, vale destacar que essa é característica que não necessariamente implica dizer que um vídeo de maior duração possui um número maior de dados para analisar. As tabelas que resumem os dados de cada vídeo no início de suas respectivas seções as quais deixam evidente essa relação diante do número de trajetos anotados no período. Os vídeos 14 do LOST e o Ped2 da UCSD possuem um tempo de observação mais curto porém são mais densos em movimentação de objetos. Devido a falta de robustez utilizada no rastreamento dos vídeos do LOST, muitos objetos tiveram associados à eles próprios, vários fragmentos curtos de trajetos que ganharam identificações diferentes nas anotações. Portanto, apesar dos vídeos do LOST serem mais longos, a seleção dos melhores trajetos durante as anotações coloca o conjunto de anotações com tamanho comparável a outros *datasets* de menor período de observação.

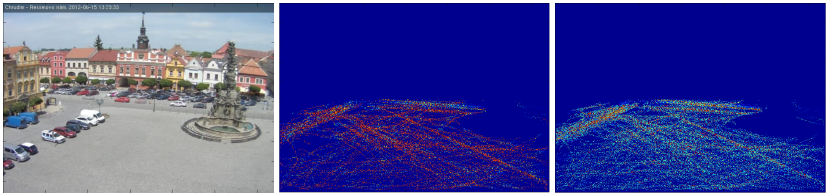
Os gráficos mostrados nas seções seguintes ilustram as simulações mais representativas das análises dos 4 vídeos selecionados para avaliar o desempenho da DMA proposta neste trabalho. Em todos eles, foram segmentadas as análises a partir do histórico de até 20 ou até 40 transições do movimento realizado por qualquer objeto. O valor $\tau = 20$ foi escolhido baseado nas considerações de que esse valor é maior do que o mínimo tamanho de trajetos rastreados e que também representa movimentos globais distintos mesmo entre regiões com pequeno fator de grade. Neste aspecto, para fatores de grade maiores, as 20 transições de um objeto pode representar no histórico do treinamento, até mais do que toda a participação de um objeto que usa toda ou quase toda a área da cena. Em contrapartida um número muito alto de τ pode levar a um custo computacional elevado pois cada transição de um objeto gera até $\frac{\tau(\tau+1)}{2}$ vetores de transição para cada região que é cruzada por um trajeto. A desvantagem do custo mencionado não foi experimentado por conta da limitação e redução de amostras proporcionada pela abordagem tra-

tada no presente trabalho. De qualquer forma, utilizou-se o valor de $\tau = 40$ como recurso para avaliar seus efeitos na capacidade do modelo aprender trajetos mais complexos, por serem mais longos.

4.1 SIMULAÇÕES E AVALIAÇÃO DO VÍDEO 1 DO LOST

A Figura 29(a) mostra o detalhe do fundo estático da cena do vídeo 1 e ao lado, nas Figuras 29(b) e 29(c) estão ilustradas as distribuições das amostras em cada *pixel* do *frame* para $\tau = 20$ e $\tau = 40$ respectivamente. Essa distribuição foi alcançada durante a primeira fase de treinamento do Algoritmo 1 para um fator de grade unitário considerando sempre o caso de maior tempo de amostragem do rastreamento que no caso desse vídeo, é de 240 minutos. Para as ilustrações dos demais vídeos nas próximas seções, a metodologia é a mesma. As Figuras 29(b) e 29(c) apresentam a normalização para o máximo de amostras encontradas em uma única região. Dessa forma, as cores mais intensas (em vermelho) indicam uma concentração maior de amostras. Em um cenário ideal, diferente de todos os vídeos avaliados, uma distribuição deveria ser homogênea em toda área da ROI. Observa-se que este vídeo possui uma distribuição relativamente equilibrada de amostras especialmente pela característica de captar imagens de uma praça pública, ocupada pela movimentação exclusiva de pedestres em grande parte da área da ROI. Uma maior concentração de amostras está associada ao longo das áreas de duas ruas e respectivo cruzamento. A câmera está configurada com um largo ângulo de abertura na captura das imagens registrando assim, objetos móveis muito distantes dela. Isso resultou em um *dataset* que inclui objetos com variados tamanhos de *bounding box*, chegando a limites inferiores de 4 *pixels* tanto em largura (w) quanto de altura (h).

Figura 29: Frame e distribuição das amostras do vídeo 1 do LOST.



(a) Fundo estático da cena. (b) $\tau = 20$. (c) $\tau = 40$.

Fonte: O *frame* médio em (a) é disponibilizado por Abrams et al. (2012).

Mais informações sobre o *dataset* estão descritas na Tabela 5 que com-

plementam as informações disponíveis na Tabelas 3 e 4 da seção 3.4. Tabelas para cada vídeo como essa bem como a distribuição das amostras, serão apresentadas também nas próximas seções para servir de referência nas análises dos resultados e conclusões do presente trabalho. Algumas das informações disponíveis nas tabelas aparecem no título dos gráficos de desempenho como por exemplo, a identificação (**ID**) do *dataset*.

Tabela 5: Dados das anotações e informações do vídeo 1 ($p_u = 1$).

Dataset(ID)	Tamanho (minutos)	Trajetos normais	Trajetos anormais	Transições	Amostras $\tau = 20$	Amostras $\tau = 40$	w/h mínimo	w/h médio	Fase 1	Fase 2
Lost 1.c4	120	601	15	29410	429137	692827	4/4	16/22	✓	✓
Lost 1.c4.f1	120	601	-	28501	429137	692827	4/4	16/22	✓	
Lost 1.c4.f2	120	589	22	26252	-	-	4/4	18/21		✓
Lost 1.c8	240	1190	37	53714	798048	-	4/4	17/22	✓	✓

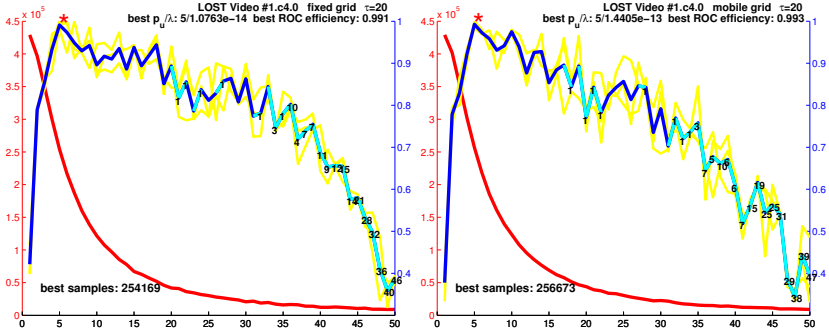
Nas duas últimas colunas da Tabela 5 há um *checkmark* que representa situações bem distintas do uso dos *datasets* nas fases de treinamento do modelo de aprendizagem. Em função do treinamento supervisionado adotado algumas variações do uso dos *datasets* foram avaliadas. O vídeo marcado exclusivamente para ser utilizado na fase 1 somente possui trajetos normais em sua base de anotações. Já os vídeos marcados como uso exclusivo na fase 2, apresentam trajetos normais e anormais diferentes daqueles utilizados na fase 1. Por fim, os *datasets* de vídeos que são marcados como fase 1 e 2, usam na fase 2, os mesmos trajetos normais usados na fase 1 juntamente com outro conjunto de trajetos anormais. Os *datasets* de vídeos que possuem ID finalizados com a notação *.f1* e *.f2* serão utilizados somente no próximo capítulo como base para testar a capacidade de generalização da aprendizagem da abordagem proposta. Embora não seja objetivo principal da tese, não se dispensou explorar e avaliar o comportamento da DMA proposta diante de outros ensaios, os quais possam servir de estímulo para trabalhos futuros.

4.1.1 Desempenho com Trajetos Até 20 Transições

As Figuras 30(a) e 30(b) ilustram o desempenho da análise do vídeo 1 do LOST considerando 4 segmentos de 30 minutos de anotações do *dataset* e janela de transições $\tau = 20$. Na sequência, as Figuras 31(a) e 31(b) mostram o desempenho com o mesmo valor τ, porém com o dobro de segmentos de vídeo. O comportamento da análise para os dois tipos de grade praticamente são similares levando a indicar que o valor de $p_u = 5$ é o agrupamento que leva a DMA à melhor eficiência no reconhecimento de movimentos anormais e normais, chegando em algumas situações, a 100% de sucessso nas

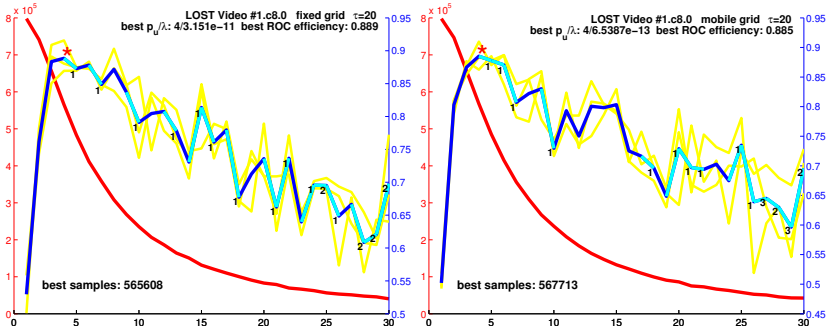
inferências. Importante destacar que essa condição foi comprovada com o modelo de teste implementado pelo Algoritmo 2, abordado no capítulo anterior. Para tanto, o valor médio de λ bem como o fator de grade foram extraídos desse treinamento como bases para testar qualquer um dos vídeos que participaram do treinamento, porém agregados com trajetos anormais. Esse procedimento se repetiu para todas as simulações.

Figura 30: Desempenho com 120 minutos do vídeo 1 e $\tau = 20$.



(a) Usando grade fixa. (b) Usando grade móvel.

Figura 31: Desempenho com 240 minutos do vídeo 1 e $\tau = 20$.



(a) Usando grade fixa. (b) Usando grade móvel.

Outro aspecto que merece destaque na comparação entre as simulações, especialmente no que se refere ao tipo de grade, é o valor de λ. Um valor maior nessa variável indica que os valores mínimos de probabilidades entre todas as transições avaliadas são maiores, e isso é um bom sinal de quali-

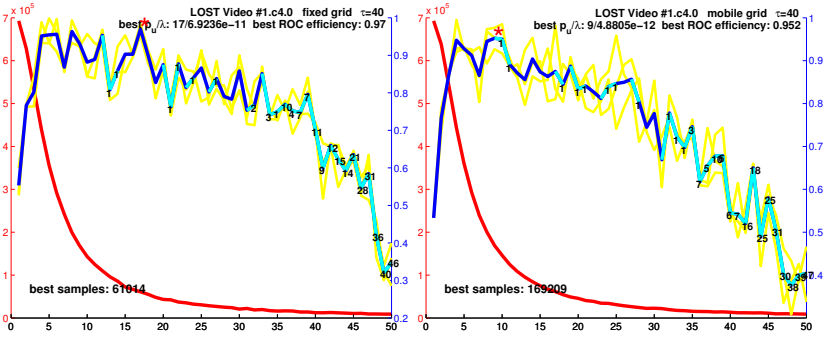
dade do modelo em função de que há uma direção no entendimento de que existem agrupamentos que reuniram conjuntos de amostras mais densas e correlatas. Tratando-se de um modelo estatístico, essa condição é oportuna pois muitos grupos de amostras apresentam pequenas variâncias entre si, formando componentes (*clusters*) que convergem com rapidez na aprendizagem estatística com EM. Nessa linha de raciocínio, observa-se que praticamente dobrar o número de amostras, não melhora o desempenho do modelo. Muito pelo contrário, além da queda de eficiência, a instabilidade da convergência do algoritmo EM que é determinante para as medidas em novas rodadas de treinamento ficam mais visíveis pelas diferentes oscilações presentes nas diferentes curvas de desempenho apresentadas em amarelo. Isso sugere que o aparecimento de mais amostras em cada região da grade as quais tornaram a disjunção entre elas cada vez mais fraca, ou seja, com baixa similaridade entre as classes (intra-classes) e alta similaridade entre classes (inter-classes). Como consequência, a cada nova rodada do EM, se ele convergir nessa situação, os resultados podem ser completamente diferentes dos anteriores. Portanto a qualidade das curvas de desempenho pode ser subjetivamente avaliada pela distorção entre a curva média e as demais curvas que participaram do levantamento dessa média.

4.1.2 Desempenho com Trajetos Até 40 Transições

As Figuras 32(a) e 32(b) ilustram o desempenho da análise do vídeo 1 do LOST considerando 4 segmentos de 30 minutos de anotações do *dataset* e janela de transições $\tau = 40$. Nestas simulações e dos demais vídeos, optou-se em dobrar a janela de observação para observar o comportamento do modelo frente a análise de movimentos mais longos e portanto mais complexos. Naturalmente que o número de amostras para os mesmos 120 minutos de vídeo praticamente dobram por conta de que cada região agrega agora, até 39 amostras de cada transição de objeto que por ela atravessa.

O modelo ainda preservou a qualidade nas inferências se comparado com a mesma situação de $\tau = 20$. Também, mesmo com um número maior de amostras, se comparado com os ensaios de 240 minutos de vídeo, esboçou uma estabilidade melhor nas curvas. A incerteza maior ficou associada pela divergência na busca do valor ideal de p_u entre os dois tipos de grade e nelas propriamente. Mesmo assim, esse ensaio demonstrou a capacidade do modelo proposto em aprender trajetos curtos e longos com similar desempenho.

Figura 32: Desempenho com 120 minutos do vídeo 1 e $\tau = 40$.



(a) Usando grade fixa. (b) Usando grade móvel.

4.1.3 Resumo e Avaliação dos Resultados

A Tabela 6 sumariza todos os principais resultados das simulações representativas do video 1 do LOST. Por essa razão as simulações relacionadas com 240 minutos de tamanho (*.c8*) e $\tau = 40$ não aparecem tabuladas. Para todas as simulações observou-se um valor ótimo de p_u que variou entre 4 e 17. Em uma avaliação preliminar isso significa que um valor de $p_u = 4$ para a resolução desse vídeo de 640x480 *pixels*, leva uma grade fixa possuir 160x120 diferentes regiões de transições. Então, o movimento de um objeto, considerando $\tau = 40$ pode ser avaliado com um histórico de transições de até 25% da largura ou até 30% da altura da cena. Se for considerado somente as regiões sobre a área da ROI, praticamente todos os movimentos globais são avaliados dentro de uma única janela de observação. Essa situação ficou mais evidenciada com o modelo baseado em grade tipo móvel em função da acomodação exclusiva de regiões somente sobre a ROI.

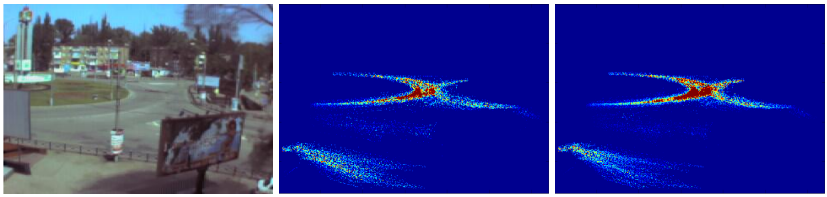
Tabela 6: Melhores resultados do *dataset* do vídeo 1 do LOST.

Janela	**Dataset(ID)**	p_u		**Amostras**		**%Amostras**		λ		**ROCefficiency**	
		fixa	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*
$\tau = 20$	Lost 1.c4	5	5	254169	256673	59,2	59,8	1,08e-14	1,44e-13	0,991	0,993
	Lost 1.c8	4	4	565608	567713	70,9	71,1	3,16e-11	6,54e-13	0,889	0,885
$\tau = 40$	Lost 1.c4	17	4	61014	169209	8,8	24,4	6,93e-11	4,88e-12	0,970	0,952

4.2 SIMULAÇÕES E AVALIAÇÃO DO VÍDEO 14 DO LOST

Da mesma forma como foi feito no vídeo 1, a Figura 33(a) mostra o detalhe do fundo estático da cena do vídeo 14 e ao lado, as Figuras 33(b) e 33(c) ilustram a distribuição das amostras em cada *pixel* do *frame* para $\tau = 20$ e $\tau = 40$ respectivamente. Diferente dos demais vídeos avaliados, esse vídeo possui uma massiva quantidade de amostras concentradas na área da ROI que está associada a uma rotatória de trânsito onde circulam diversos tipos de automóveis identificados adequadamente conforme a Tabela 1 da subseção 3.1.3. Em outra área bem distinta da ROI há uma estação de metrô por onde circulam pedestres de forma moderada, povoando a região com amostras mais distribuídas. Uma área de oclusão devido a uma grande placa publicitária interrompe o rastreamento dos objetos que saem e entram na rotatória e que tem origem e destino pelo lado direito inferior do *frame*. Em função dessas características, existem vários trajetos curtos e outros tantos que desenvolvem o mesmo tipo de trajeto e com poucas variações de velocidade como é o caso daqueles que estão circulando na rotatória. Mais detalhes sobre esse *dataset* estão descritos na Tabela 7.

Figura 33: Frame e distribuição das amostras do vídeo 14 do LOST.



(a) Fundo estático da cena. (b) $\tau = 20$. (c) $\tau = 40$.

Fonte: O *frame* médio em (a) é disponibilizado por Abrams et al. (2012).

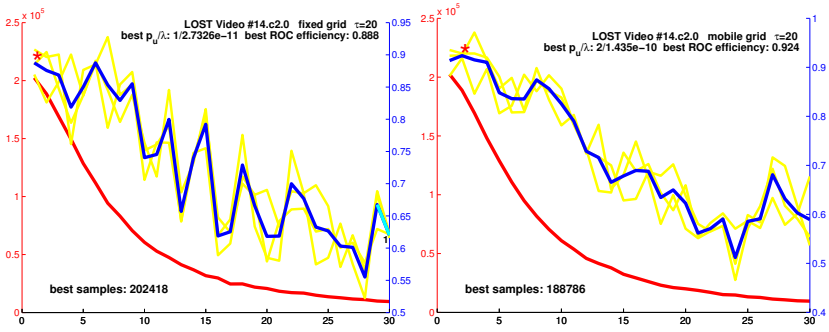
Tabela 7: Dados das anotações e informações do vídeo 14 ($p_u = 1$).

Dataset(ID)	**Tamanho**	**Trajetos**		**Transições**	**Amostras**		**w/h**		**Fase**	
	(minutos)	**normais**	**anormais**		$\tau = 20$	$\tau = 40$	**mínimo**	**médio**	**1**	**2**
Lost 14.c2	60	516	17	15699	202418	268523	8/8	22/19	✓	✓
Lost 14.c4	120	891	22	29540	381591	519417	8/8	23/19	✓	✓

4.2.1 Desempenho com Trajetos Até 20 Transições

As particularidades da repetição dos mesmos trajetos que são amostrados na segunda maior taxa de FPS (8.9 em média) dentre os demais vídeos, levou a uma limitação na seleção de amostras durante as anotações desse vídeo para evitar excessos de amostras repetidas. Isso se fez necessário para evitar os problemas de convergência do algoritmo EM diante de amostras repetidas ou muito similares. Além da limitação da seleção de amostras, também se fez uso de tempos menores de amostragem de 30 e 60 minutos que foram suficientes para análise desse cenário. As Figuras 34(a) e 34(b) ilustram o desempenho da análise do vídeo 14 do LOST considerando 2 segmentos de 30 minutos de anotações do *dataset* e janela de transições $\tau = 20$.

Figura 34: Desempenho com 60 minutos do vídeo 14 e $\tau = 20$.

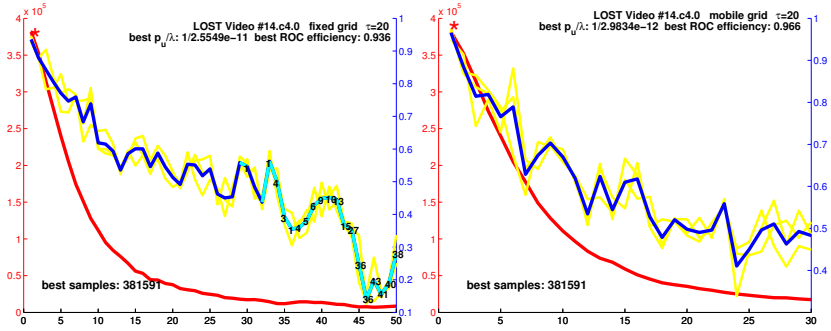


(a) Usando grade fixa. (b) Usando grade móvel.

As curvas mostram claramente que o melhor desempenho já ocorre com fator de grade unitário ou igual a 2 no caso da grade móvel. Conclui-se que, em nível de *pixel*, já se produziram amostras no mínimo suficientes para extrair o melhor resultado da DMA. Outras simulações não representadas aqui conduziram ao mesmo comportamento para vídeos acima de 120 minutos. Em todos os casos o melhor desempenho ocorreu quando $p_u = 1$ e com resultados piores de *ROCefficiency* quanto maior o tempo de vídeo amostrado. Destaca-se que para o caso das curvas da Figura 34 o arranjo das regiões uniformes da grade móvel proporcionaram uma qualidade maior nos agrupamentos das amostras e uma consequente estabilidade nos resultados de novas rodadas de treinamento. Além disso na no caso de 60 minutos a grade móvel apresentou o melhor resultado e $p_u > 1$, indicando que o limite de amostragem para este cenário é de fato menor que 60 minutos. Para con-

solidar as observações acima, as Figuras 35(a) e 35(b) ilustram o desempenho da análise do vídeo 14 do LOST considerando 4 segmentos de 30 minutos de anotações do *dataset* e janela de transições $\tau = 20$.

Figura 35: Desempenho com 120 minutos do vídeo 14 e $\tau = 20$.



(a) Usando grade fixa. (b) Usando grade móvel.

Diante dos resultados previstos para o período de 120 minutos indicando $p_u = 1$ como o melhor fator de grade, fica dispensável detalhar a mesma simulação com $\tau = 40$ que também convergiu para o mesmo fator unitário de grade. O destaque fica por conta da grade móvel que novamente esboçou melhor rendimento.

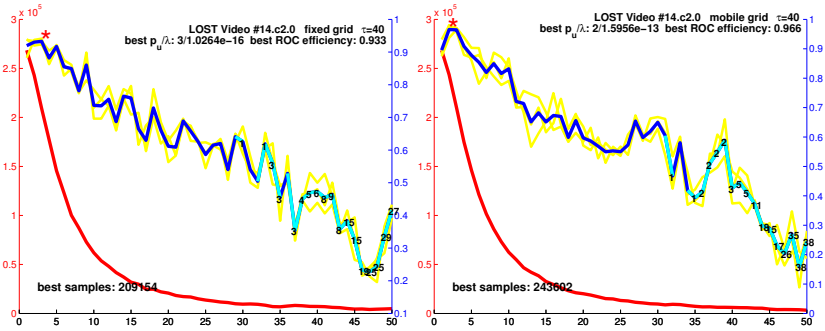
4.2.2 Desempenho com Trajetos Até 40 Transições

As Figuras 36(a) e 36(b) ilustram o desempenho da análise do vídeo 14 do LOST considerando 2 segmentos de 30 minutos de anotações do *dataset* e janela de transições $\tau = 40$.

4.2.3 Resumo e Avaliação dos Resultados

A Tabela 8 resume todos os principais resultados das simulações representativas do vídeo 14 do LOST. A concentração de trajetos em uma região limitada da ROI como no caso do vídeo 14, causa uma super-amostragem (*oversampling*) que satura rapidamente o modelo, prejudicando mais do que ajudando na análise. Soluções para o *oversampling* podem ser resolvidas com técnicas de regressão estatística (BISHOP, 2006) ou ainda com critérios heurísticos de parada e atualização periódica do conjunto de amostras para

Figura 36: Desempenho com 60 minutos do vídeo 14 e $\tau = 40$.



(a) Usando grade fixa. (b) Usando grade móvel.

cada região. No contexto deste trabalho, esta última solução parece ser mais factível frente ao objetivo de se se reduzir esforço computacional, privilegiando aplicações no mundo real.

Tabela 8: Melhores resultados do *dataset* do vídeo 14 do LOST.

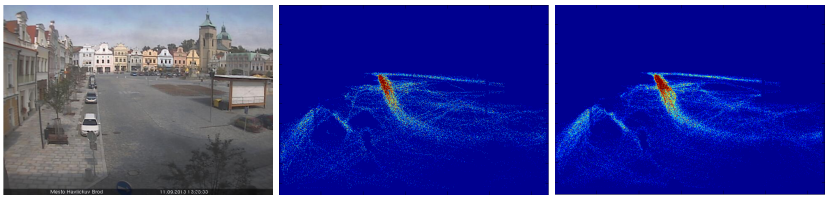
Janela	**Dataset(ID)**	p_u		**Amostras**		**%Amostras**		λ		**ROCefficiency**	
		fixa	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*
$\tau = 20$	Lost 14.c2	1	2	202418	188786	100,0	93,3	2,73e-11	1,44e-10	0,888	0,924
	Lost 14.c4	1	1	381591	381591	100,0	100,0	2,56e-11	2,99e-12	0,936	0,966
$\tau = 40$	Lost 14.c2	3	2	209154	243602	77,9	90,7	1,03e-16	1,60e-13	0,933	0,966

4.3 SIMULAÇÕES E AVALIAÇÃO DO VÍDEO 17 DO LOST

A Figura 37(a) mostra o detalhe do fundo estático da cena do vídeo 17 e ao lado, as Figuras 37(b) e 37(c) ilustram a distribuição das amostras em cada *pixel* do *frame* para $\tau = 20$ e $\tau = 40$ respectivamente. Esse vídeo possui características similares as do vídeo 1 no que se refere ao ângulo de abertura da câmera e grande áreas públicas utilizadas por pedestres na calçada à esquerda do *frame* e na praça pública à direita. O cenário é atravessado por uma rua que possui um trafego moderado de automóveis e que portanto possui grande concentração e representatividade de amostras em situação parecida com a encontrada no vídeo 14.

Mais detalhes sobre esse *dataset* estão descritos na Tabela 9. Tal como no vídeo 1, as avaliações com $\tau = 40$ e 240 minutos contribuíram pouco para análise e por isso não foram tabuladas e relatadas.

Figura 37: Frame e distribuição das amostras do vídeo 17 do LOST.



(a) Fundo estático da cena. (b) $\tau = 20$. (c) $\tau = 40$.

Fonte: O *frame* médio em (a) é disponibilizado por Abrams et al. (2012).

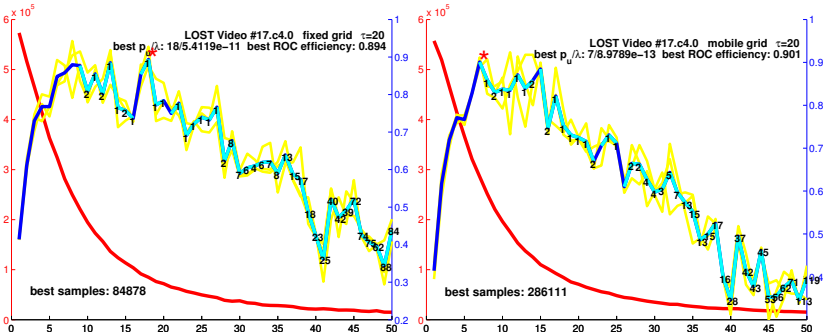
Tabela 9: Dados das anotações e informações do vídeo 17 ($p_u = 1$).

Dataset(ID)	Tamanho (minutos)	Trajetos normais	Trajetos anormais	Transições	Amostras $\tau = 20$	Amostras $\tau = 40$	w/h mínimo	w/h médio	Fase 1	Fase 2
Lost 17.c4	120	1171	31	41896	557365	831147	5/5	24/31	✓	✓
Lost 17.c4.f1	120	1171	-	40594	557365	831147	5/5	24/31	✓	
Lost 17.c4.f2	120	1243	63	51053	-	-	5/5	25/32		✓
Lost 17.c8	240	2414	94	89366	1253996	-	5/5	26/33	✓	✓

4.3.1 Desempenho com Trajetos Até 20 Transições

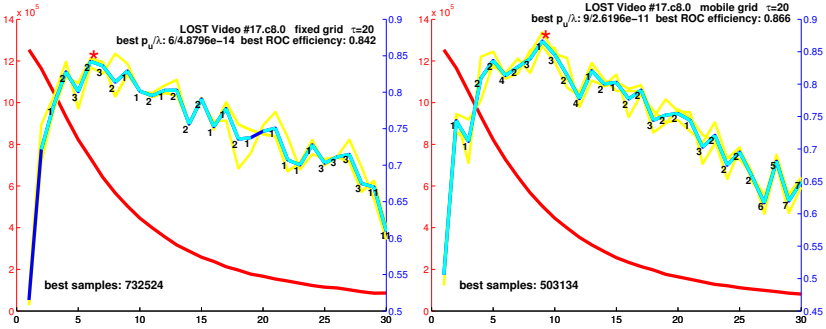
As Figuras 38(a) e 38(b) ilustram o desempenho da análise do vídeo 17 do LOST considerando 4 segmentos de 30 minutos de anotações do *dataset* e janela de transições $\tau = 20$. Na sequência, as Figuras 39(a) e 39(b) mostram o desempenho com o mesmo valor τ, porém com o dobro de segmentos.

Figura 38: Desempenho com 120 minutos do vídeo 17 e $\tau = 20$.



(a) Usando grade fixa. (b) Usando grade móvel.

Figura 39: Desempenho com 240 minutos do vídeo 17 e $\tau = 20$.



(a) Usando grade fixa. (b) Usando grade móvel.

Em linhas gerais, a redução de amostras e consequente eficiência tornar-se-á mais relevante quando o valor de p_u ficar próximo a média das dimensões dos objetos características de cada cenário. As curvas do vídeo 17, especialmente aquelas da grade móvel na Figura 38(b), revelam uma região onde o desempenho, antes de entrar em queda livre, flutua em torno de 10% do valor médio de *ROCefficiency* desde o melhor valor de p_u até quando ele chega próximo ao valor médio de largura dos objetos deste cenário, que é 24, conforme a Tabela 9. A maior parte da movimentação dos objetos neste cenário é na direção diagonal ou horizontal, justificando o porque o valor próximo a 24 é o ponto que marca a queda mais importante da eficiência da DMA.

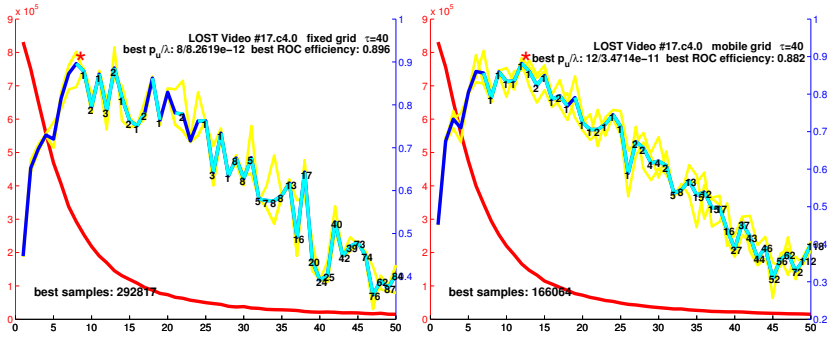
4.3.2 Desempenho com Trajetos Até 40 Transições

As Figuras 40(a) e 40(b) ilustram o desempenho da análise do vídeo 17 do LOST considerando 4 segmentos de 30 minutos de anotações do *dataset* e janela de transições $\tau = 40$.

4.3.3 Resumo e Avaliação dos Resultados

A Tabela 10 sumariza todos os principais resultados das simulações representativas do video 17 do LOST. O comportamento da análise do vídeo 17 não diferencia muito do que foi avaliado até aqui, exceto pelo fato de que ocorreram períodos maiores que 60 minutos do vídeo 14. Observa-se sempre

Figura 40: Desempenho com 120 minutos do vídeo 17 e $\tau = 40$.



(a) Usando grade fixa. (b) Usando grade móvel.

um aclive acentuado na eficiência da DMA quando a área das regiões da grade começa a aumentar até que ela atinge um valor máximo. A partir desse ponto, como já discutido anteriormente, a redução da eficiência é mais lenta e vai depender do tipo de cenário principalmente por conta da redução gradativa do número de amostras. Esta dependência está ligada as dimensões de largura e altura do *bounding box* dos objetos que produziram as amostras. Nesse contexto, de acordo com a metodologia usada no modelo de movimento da DMA, se a área da região da grade se tornar maior do que as dimensões da maioria dos objetos, haverá uma queda significativa de amostras que representam aquela movimentação inócua dentro cada região.

Tabela 10: Melhores resultados do *dataset* do vídeo 17 do LOST.

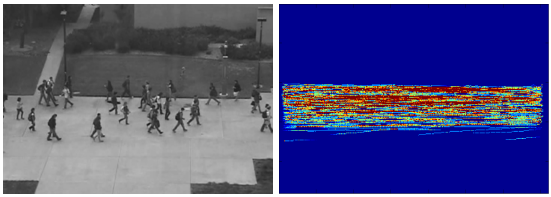
Janela	**Dataset(ID)**	p_u		**Amostras**		**%Amostras**		λ		**ROCefficiency**	
		fixa	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*
$\tau = 20$	Lost 17.c4	18	7	84878	286111	15,2	51,3	5,41e-11	8,98e-13	0,894	0,901
	Lost 17.c8	6	9	732524	503134	58,4	40,1	4,88e-14	2,62e-11	0,842	0,866
$\tau = 40$	Lost 17.c4	8	12	292817	166064	35,2	20,0	8,26e-12	3,47e-11	0,896	0,882

4.4 SIMULAÇÕES E AVALIAÇÃO DO VÍDEO PED2 DA UCSD

A Figura 41(a) mostra um exemplo de *frame* do vídeo Ped2 e ao lado, na Figura 41(b) está a distribuição das amostras em cada *pixel* do *frame* para $\tau = 40$. Este vídeo tem características completamente diferentes de todos os da base do LOST. Ele é um *dataset* comumente utilizado por outros auto-

res visando a realização de trabalhos voltados para abordagens baseadas em movimento e análise baseada em *pixel* ou região, incluindo análise de multidão devido a grande densidade de objetos presentes em muitas sequências de *frames*.

Figura 41: Frame e distribuição das amostras do vídeo Ped2 da UCSD.



(a) Um exemplo de *frame*. (b) $\tau = 40$.

A área da ROI é limitada quase ao centro do *frame* e o movimento intenso de pedestres nas duas direções tornam bastante denso o número de amostras nessa área. O ângulo de abertura da câmera é mais fechado, levando o *bounding box* dos objetos com uma média alta de $15/31 pixels$ de w/h. Muitos trajetos são curtos devido a média de tempo de reprodução de 4 segundos de cada sequência de *frames*. Para as sequências de teste, são utilizados *frames* que possuem movimentação de pessoas com bicicleta, skate ou mesmo pequenos veículos sobre a área exclusiva de uso de pedestres. Este *dataset* foi escolhido propositalmente para avaliar o comportamento da DMA proposta usando sequências muito curtas de vídeos e com alta densidade de objetos. Devido a particularidade dos trajetos breves e normalmente retilíneos e em direções opostas, sem qualquer complexidade, a análise demonstrada aqui ficou restrita ao desempenho com grade fixa pois a grade móvel, posicionada sobre um conjunto de amostras dispostas exclusivamente em uma faixa horizontal, conforme detalhado na Figura 41(b), não contribui significativamente na avaliação. Mais detalhes sobre esse *dataset* estão descritos na Tabela 11.

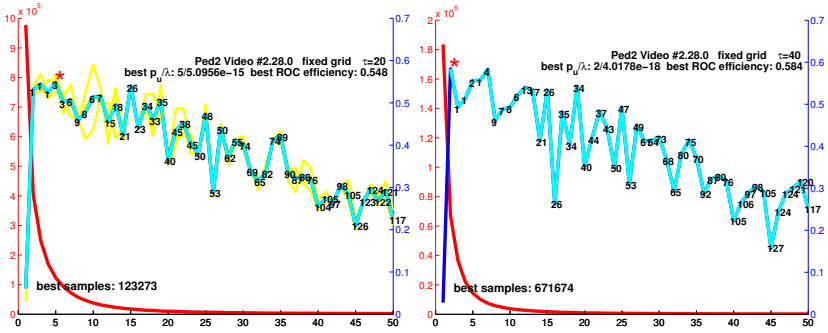
Tabela 11: Dados das anotações e informações do vídeo Ped2 ($p_u = 1$).

Dataset(ID)	**Tamanho**	**Trajetos**		**Transições**	**Amostras**		**w/h**		**Fase**	
	(minutos)	**normais**	**anormais**		$\tau = 20$	$\tau = 40$	**mínimo**	**médio**	**1**	**2**
Ped2	2,5	447	26	55896	977256	1835771	4/25	16/33	✓	✓
Ped2.f1	1,5	265	-	31626	550676	-	4/25	16/33	✓	
Ped2.f2	1,0	182	25	24008	550676	-	6/16	17/32		✓

4.4.1 Desempenho com Trajetos Até 20 e 40 Transições

A Figura 42 ilustra o desempenho da análise do vídeo Ped2 considerando toda a sequência de *frames* de treinamento e teste.

Figura 42: Desempenho com 2,5 minutos do vídeo Ped2 em grade fixa.



(a) Com $\tau = 20$. (b) Com $\tau = 40$.

4.4.2 Resumo e Avaliação dos Resultados

A Tabela 12 resume os principais resultados da simulação do vídeo Ped2. Apesar da maior quantidade de amostras entre todos os *datasets*, o Ped2 resultou nas piores curvas de desempenho. Uma combinação do alto valor de FPS, tipo de cenário predominantemente formado por pedestres e período muito curto de amostragem implicou na rápida queda de número de amostras quando o valor de p_u é incrementado. Mesmo assim houve o pico de melhor desempenho, comum em todos os ensaios, quando o fator de grade alcançou o valor que garantiu o melhor desempenho da DMA. Mesmo sendo um *dataset* formado por anotações fieis de vídeo, pressupondo assim um rastreamento robusto, o curto período de amostragem não permitiu acumular amostras suficientes para reconhecer trajetos mais longos. A simulação com esse *dataset* foi útil para concluir que a abordagem da DMA proposta no presente trabalho só é eficiente para *datasets* exclusivamente de longa duração típicas das aplicações de videovigilância. No entanto, vale destacar que o baixo desempenho não está associado a escala de tempo, mas sim dela com o tamanho dos objetos observados. Assim, a DMA só seria eficiente se os vários FPS implicassem em movimentações rápidas dos objetos onde, a cada

frame, estivessem em locais da cena em distancias acima de pelo menos o seu próprio tamanho em altura ou largura. No caso do Ped2, 30 *frames* possuem o registro de 2 ou 3 passos de uma pessoa em ritmo normal de caminhada, implicando em um número reduzido de amostras para o treinamento do GMM. Por esse motivo, as quase 1 milhão de amostras geradas das anotações de *frames* quando $\tau = 20$ são reduzidas para menos da metade quando $p_u = 2$ pois no tempo de $1/30s$, grande parte das amostras ainda se encontram dentro da pequena região de $2x2pixels$. Mesmo diante dessa adversidade, o modelo proposto aqui cumpriu seu papel, encontrando os melhores arranjos de grade fixa para os diferentes valores de τ.

Tabela 12: Resultados do *dataset* do vídeo Ped2.

Janela	**Dataset(ID)**	p_u		**Amostras**		**%Amostras**		λ		**ROCefficiency**	
		fixa	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*
$\tau = 20$	Ped2	5	-	123273	-	12,6	-	5,10e-15	-	0,548	-
$\tau = 40$	Ped2	2	-	671674	-	36,6	-	4,02e-18	-	0,584	-

4.5 CONSIDERAÇÕES SOBRE OS RESULTADOS

Comportamentos comuns foram observados em todas as simulações. Conforme mencionado na avaliação do vídeo 17 do LOST, o desempenho sempre começa ruim para o fator de grade unitário mesmo com a maior quantidade de amostras disponíveis. Isso porque a posição do centroide de um objeto pode estar em diversas possibilidades de regiões (*pixels*) nas próximas transições. Como consequência, o cálculo das probabilidades mínimas vai resultar em valores nulos ou muito pequenos ou ainda eventualmente altos, dificultando a busca por um limiar λ que consiga ser eficiente na classificação do movimento. A medida que o valor de p_u sobe, essas dificuldades diminuem e a relação entre os trajetos, as amostras e as regiões vai ficando mais coerente até o ponto onde as amostras, mesmo em menor número, conseguem representar bem os movimentos globais. A partir desse ponto, principalmente pela redução continuada do número de amostras, o valor de *ROCefficiency* vai caindo sistematicamente.

Perante a diversidade de cenários envolvidos nas simulações, a DMA proposta demonstra possuir uma insensibilidade ao contexto da cena e da falta de robustez do rastreamento dos objetos. Existe uma semelhança nas curvas de desempenho para todos os tipos de vídeos analisados mesmo eles possuindo diferentes resoluções, FPSs, ROIs, qualidade do rastreamento, tamanho dos trajetos e quantidade de trajetos anormais. Essa independência é muito bem-vinda para aplicações voltadas à videovigilância, foco de atenção

do presente trabalho.

4.5.1 Quantidade de Amostras

Para entender o quanto o aumento da quantidade de amostras influencia na eficiência da DMA, foram aproximadamente dobrados os valores de amostras na análise dos vídeos 1 e 17 do LOST (ID 1.c8 e 17.c8 respectivamente). Os resultados mostraram que além da convergência para o um valor de p_u muito próximo daquele com metade das amostras, a eficiência da DMA piorou. Essa mesma linha de raciocínio ocorreu na avaliação do vídeo 14 do LOST amostrado com 120 minutos, o qual produziu valores de p_u ideal unitários. Esses ensaios demonstram a situação de *oversampling* que implica no *overtraining* do GMM, viciando o modelo a reconhecer os exemplos dos dados normais, errando mais a classe minoritária dos trajetos anormais.

Em um comparativo com a proposta dos autores (BASHARAT et al., 2008) que inspirou o modelo de movimento proposto aqui, fica claro que a qualidade de um DMA não está ligado com o número de amostras, mas sim com a seletividade delas. No caso, os autores perceberam que somente a captura dos vetores de transição da posição do centroide dos objetos tornava a modelagem de movimento espacialmente esparsa. Para resolver isso eles atualizaram todos os *pixels* (regiões) até o limite da área do *bounding box* de todos os objetos que criaram os 1342 tracks usados na fase de treinamento. Essa densidade quadraticamente maior de amostras, ajudou a reduzir fontes de erro no modelo de cena mas criou restrições computacionais, uma vez que a quantidade de amostras ficou dependente da quantidade de transições que o objeto realiza ao longo dos *frames*. Os autores fixaram o limite $\tau = 20$. Isso resultou no uso de mais de 250 vezes o número de amostras utilizadas em qualquer vídeo avaliado no presente trabalho. Essa enorme diferença se deve ao fato de que o modelo de movimento desses autores faz cópias de amostras em todos os *pixels* da área de cada *bounding box*. Em uma simulação, usando o conjunto de anotações do vídeo de treinamento disponibilizados pelos autores, com resolução de vídeo de 240x320 *pixels* e $\sim$3 horas de duração, o modelo de movimento proposto aqui para um fator de grade unitário, gerou mais de 250 milhões de amostras quando se utilizou a opção de reproduzir as cópias dos vetores de transição na vizinhança. O desempenho apresentado pelos autores através de uma curva ROC reproduziu um perfil conservador dado a todo esforço computacional utilizado.

Um outro cenário experimentado para verificar os efeitos do *overtraining* sobre os *datasets* avaliados aqui, foi manter a mesma quantidade de amostras para toda a variação de p_u. Como resultado, o tempo para a con-

vergência, quando ocorria, a cada rodada do EM em cada região da grade foi em média cinco vezes maior e novamente piorando a qualidade de inferências do modelo. Nesse caso, as amostras relativas ao mesmo objeto em transição na mesma região, produziam vetores redundantes onde só o valor da variável tempo era alterada sem contribuir no treinamento dos parâmetros do *pdf* de cada região.

O equilíbrio da distribuição das amostras como no caso do vídeo 1 do LOST, proporcionou um desempenho acima dos 90%, chegando em algumas simulações a 100% de eficácia nas inferências. No pior caso, com $p_u = 4$, em torno de 60% das amostras originais foram suficientes para alcançar o melhor desempenho entre todos os ensaios realizados. Os melhores deles foram alcançados com a grade móvel.

O comportamento observado da DMA sugere que é possível encontrar um número de amostras para qualquer fator de grade o qual garante eficiência máxima nas inferências. No entanto a busca dessa relação estaria na contramão dos objetivos do presente trabalho, principalmente para valores inferiores ao fator de grade ideal encontrado para cada cenário. Valores acima do fator de grade ideal continuariam demandando um número cada vez menor de amostras mas descaracterizariam a realidade do reconhecimento de padrões desejado de movimento de todos os objetos de interesse em um cenário, especialmente para objetos com dimensões pequenas.

4.5.2 Tipo de Grade no Modelo da Cena

O modelo com grade móvel proporcionou uma medida de eficiência muito próxima ou ligeiramente maior do que com a grade fixa, na maioria dos ensaios, flutuando em torno de 4% para mais no caso do vídeo 14, ou 1,8% para menos no caso do vídeo 1. Observa-se também que as oscilações das curvas de desempenho também são mais amenas na comparação. A qualidade dos arranjos de amostras em cada região da grade vai determinar o sucesso nos quesitos de similaridades para convergência do EM. Pelos resultados dos experimentos realizados até aqui, a grade móvel parece desempenhar melhor esse papel. O encapsulamento aparentemente mais adequado tanto em quantidade quanto em similaridades também foi outra consequência do uso desse tipo de grade. Em adição, a acomodação das regiões sobre a ROI proporcionada pela grade móvel elevou os valores de λ, indicando que os agrupamentos reuniram amostras mais correlatas e dessa forma tornou valores de probabilidades mais altos e portanto mais seletivos para trajetos normais. O valor de λ maior implica nessa melhora baseado no fato de que há um número maior de *clusters* encontrados durante o treinamento do algoritmo EM onde cada um

deles representa uma distribuição gaussiana relativa as amostras acumuladas sobre o trânsito de qualquer objeto que passa pela região.

Um dos fatores que contribui com o sobe e desce de valores de eficiência está relacionado com o efeito zigue zague apresentado na subseção 3.2.1.1. Além do uso da média entre várias rodadas do modelo de aprendizagem, a grade móvel também contribuiu para a atenuação desse efeito. Alguns experimentos foram realizados usando a ideia da contenção de amostragem de 2 e 3 *frames*, apresentada na mesma seção. No entanto, como era de se esperar, a redução da quantidade de amostras originais acabou afetando significativamente a análise de desempenho buscada nas simulações e por esse motivo, a contenção de amostragem só seria interessante se houvesse uma compensação de amostragem, o que não seria conveniente para os objetivos do trabalho. Independente do que elas representam, as oscilações observadas não afetam a interpretação da análise uma vez que as variações mais relevantes encontram-se na parte descendente das curvas.

4.5.3 Tamanho da Janela de Transições

Observa-se também uma taxa similar de desempenho para a análise de até 20 ou 40 transições, demonstrando a estabilidade do modelo no que se refere a capacidade de detectar anomalias em movimentos globais de trajetos de diferentes tamanhos ou complexidades. As consequências do uso de janelas maiores poderia ser melhor observado caso os vídeos avaliados possuíssem um maior número de trajetos sem fragmentação, como ocorre nas bases dos vídeos investigados aqui. Por esse motivo, tomou-se como base a análise feita com de 20 transições e como opcional a análise com o dobro dessa janela.

4.5.4 Faixa de Tamanhos Ótimos de Agrupamento

Na avaliação dos melhores valores de p_u observados em todas as simulações, observou-se uma regularidade desses com os valores relacionados as dimensões de largura ou de altura dos objetos que participaram nas fases de treinamento. Os valores citados estão informados nas Tabelas 5, 7, 9 e 11 de cada *dataset*.

O melhor valor de p_u para a maior parte das simulações convergiu para um número baixo, de 2 a 5, mas mesmo assim já representou uma redução expressiva das amostras originais da ordem de 40% em média. Isso implica em dizer que este subconjunto de amostras é o que melhor representa o cenário dentro da proposta do modelo. Também, considerando uma margem de cerca

de 10% abaixo do melhor valor de *ROCefficiency*, é possível identificar uma faixa de valores de p_u que se estende de 2 até um valor próximo do valor médio da menor dimensão do *bounding box* dos objetos. Os vídeos 1 e 17 do LOST, para ensaios com 120 minutos de treinamento mostram com mais evidência essa relação. Isso significa dizer que, em uma aplicação real, qualquer valor entre 2 e o valor médio da largura ou da altura do objeto garante um desempenho aceitável da DMA. Essa relação permite concluir que, conhecendo as dimensões médias dos objetos que se pretende monitorar, pode-se usar a grade móvel ou fixa já configurada com um valor nessa faixa e a partir dela realizar a amostragem, a aplicação do modelo de movimento e a aprendizagem.

Antes de partir para a discussão e conclusão deste trabalho, outras simulações adicionais e comparações relacionadas com a proposta serão apresentadas no próximo capitulo. Dele pretende-se somar mais alguns subsídios para consolidar a validade e aplicabilidade do tema proposto.

5 AMPLIAÇÃO DO USO DA ABORDAGEM E APLICABILIDADE EM VIDEOVIGILÂNCIA REAL

Todos os modelos desenvolvidos até aqui consideraram as condições mais ideais possíveis com o intuito de isolar as influencias das imperfeições do próprio modelo de aprendizagem proposto. Como mencionado anteriormente, foi esse o motivo que levou a adoção do método de teste com taxa de erro aparente na segunda fase de treinamento do modelo de aprendizagem. Esta estratégia permitiu extrair muitas informações dos comportamentos resultantes das simulações realizadas no capítulo anterior. Obviamente que o mundo real da videovigilância vai exigir da automação mais do que uma resposta "aparente" dos sistemas para que os mesmos se tornem úteis. Mediante os bons resultados alcançados do DMA proposto, seria uma negligência não apresentar outros resultados que considerem agora condições mais próximas da realidade. As próximas seções trazem mais alguns desses resultados bem como possíveis situações de uso da abordagem proposta.

5.1 CAPACIDADE DE GENERALIZAÇÃO DO MODELO DE APRENDIZAGEM

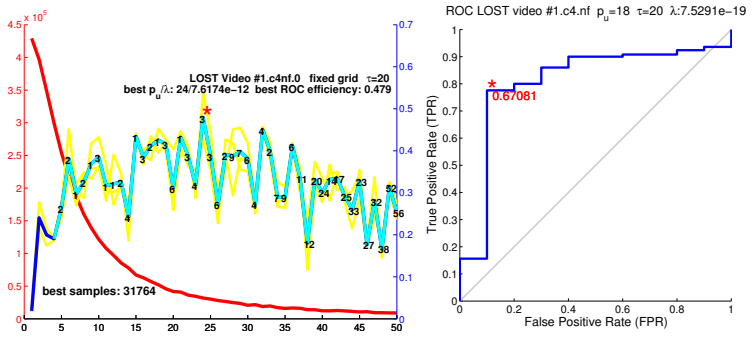
Uma das metas buscadas em reconhecimento de padrões é planejar na construção do modelo a capacidade de reconhecer padrões ainda não vistos e que não participaram do treinamento. Na DMA, esta capacidade pode ser medida pela quantidade de inferências bem sucedidas na classificação de movimentos normais e anormais quando se apresenta um novo conjunto de exemplos. Por esse motivo, muitos *datasets* disponibilizam conjuntos de exemplos específicos para o treinamento e outro diferente para testes. A modalidade de teste baseada em taxa de erro verdadeira é a forma de testar a capacidade de generalização onde a curva ROC normalmente é utilizada para avaliar o correspondente desempenho do modelo.

Pensando nisso, além do *dataset* padrão Ped2 da UCSD, que já disponibiliza conjuntos distintos de *frames* de treinamento e teste, foi criado da mesma forma, conjuntos adicionais de anotações de vídeo de teste para os vídeos 1 e 17 do LOST. Os detalhes dos dados correspondentes a essas sequencias estão disponíveis nas Tabelas 5 e 9 respectivamente, sob a indicação *.f2* no ID de cada vídeo. Para estas simulações optou-se em utilizar a mesma quantidade de segmentos de vídeos de teste que foi usada no treinamento.

5.1.1 Desempenho do Vídeo 1 do LOST

O Algoritmo 1 nessa estratégia vai usar na fase 1 do treinamento, o *frames* de vídeo sem trajetos anormais indicado na Tabela 5 com *.f1* e na segunda fase, uma nova sequência de *frames* de vídeo com trajetos normais e anormais diferentes da fase 1 são usados. O resultado da curva média de desempenho é mostrado na Figura 43(a). A curva ROC mostrada na Figura 43(b) representa a melhor curva e métrica *ROCefficiency* de uma das curvas que fez parte no cômputo da curva média de desempenho. Por esse motivo os valores de λ e da métrica possuem as divergências entre as curvas de desempenho e ROC.

Figura 43: Desempenho do vídeo 1 em taxa de erro verdadeira.



(a) Curvas de desempenho. (b) Melhor curva ROC.

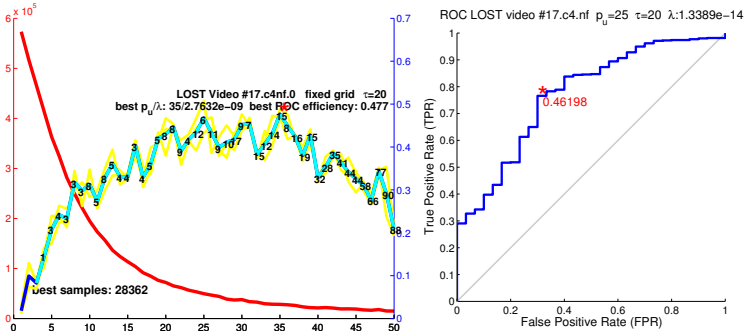
A forma de traçado da curva de desempenho se repete como aconteceu nas rodadas do capítulo anterior. No entanto observa-se uma estabilidade em torno do valor de p_u ideal que se estende de uma faixa que começa em $p_u = 6$ e se estende até próximo do valor da largura média w dos objetos pertencentes ao *dataset*. No ponto ótimo global da curva, quando $p_u = 18$, o classificador sintonizado para um valor de $\lambda = 7,5291e-19$ consegue acertar em torno de 78% dos trajetos rotulados como anormais mas erra cerca de 10% das inferências sobre trajetos normais de acordo com a curva ROC na Figura 43(b). Considerando o desempenho em geral no reconhecimento de padrões quando se testa a capacidade de generalização de um modelo, os valores alcançados estão aproximadamente na linha do que Powers (2011) considera como *good*, mesmo usando somente pouco mais de 10% das amostras iniciais. Em todas as simulações incluindo esta, é notório que a distribuição mais uniforme em toda a área da ROI, característico deste *dataset*, contribui fortemente na performance da abordagem. Isso implica dizer que em uma aplicação do mundo

real a coleta das amostras deve ser controlada segundo a densidade de movimentos que ocorrem no cenário monitorado. Ou seja, a coleta periódica deve ser mais espaçada temporalmente em locais de grande movimento tanto de pedestres quanto de veículos.

5.1.2 Desempenho do Vídeo 17 do LOST

O mesmo processo se repete para o vídeo 17. O resultado da curva média de desempenho é mostrado na figura 44(a). A curva ROC mostrada na Figura 44(b) representa a melhor curva e métrica *ROCefficiency*.

Figura 44: Desempenho do vídeo 17 em taxa de erro verdadeira.



(a) Curvas de desempenho. (b) Melhor curva ROC.

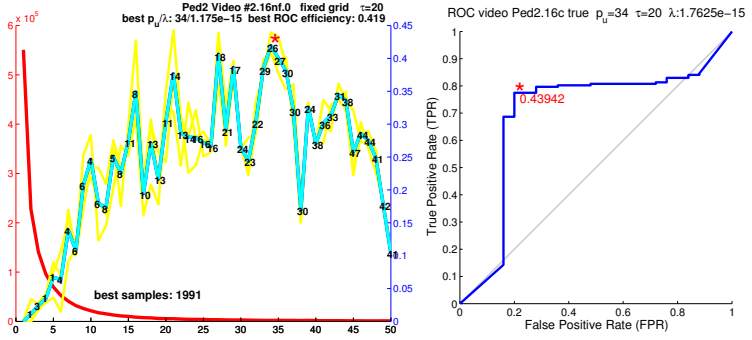
Novamente o melhor desempenho ocorre quando p_u alcança aproximadamente o valor médio das dimensões do *bounding box* dos objetos pertencentes a este *dataset*. Em função da diversidade de concentração de densidades de movimentos na ROI o desempenho da abordagem fica sofrível pois consegue acertar em torno de 78% dos trajetos rotulados como anormais porém erra cerca de 32% das inferências sobre trajetos normais. Ainda assim o modelo de aprendizagem esboça seu esforço na generalização do aprendizado.

5.1.3 Desempenho do Vídeo Ped2 da UCSD

Para concluir, mesmo não tendo um bom desempenho nas simulações do capítulo anterior, repetiu-se o processo para o *dataset* Ped2. O resultado

da curva média de desempenho é mostrado na figura 45(a). A curva ROC mostrada na Figura 45(b) representa a melhor curva e métrica *ROCefficiency*.

Figura 45: Desempenho do vídeo Ped2 em taxa de erro verdadeira.



(a) Curvas de desempenho. (b) Melhor curva ROC.

Como mencionado no capítulo anterior, devido a alta taxa de FPS, o tamanho extremamente curto de vídeo, e a densidade de movimento concentrada em toda a ROI, este *dataset* não se acomoda tão bem a uma abordagem baseada em rastreamento quanto aos demais avaliados. No entanto o comportamento deste vídeo frente a sequências diferentes na fase de teste, praticamente manteve o mesmo traçado dos vídeos mais longos anteriores, porém com um desempenho pior. Isso demonstra que a representatividade de somente 2 mil amostras espalhadas pela ROI foram suficientes para aprender a maioria dos trajetos curtos, os quais basicamente se repetem ao longo dos *frames* de teste.

A Tabela 13 resume os principais resultados das simulações dos dois vídeos do LOST e do vídeo Ped2 da UCSD. Eles são suficientes para fins comparativos com os ensaios realizados com taxa de erro aparente do capítulo anterior. A notação *.f* na composição do ID do vídeo é somente para diferenciar os resultados entre os modos de treinamento.

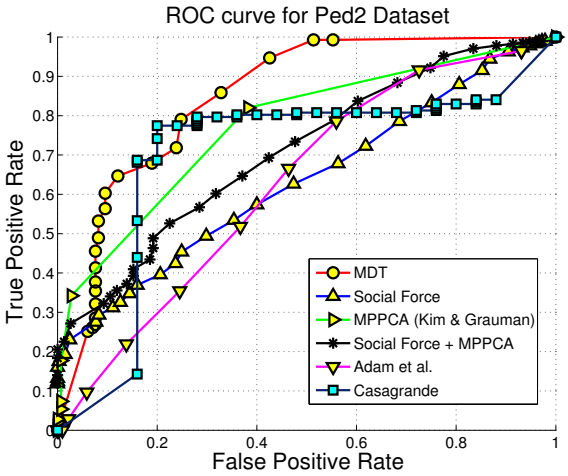
Tabela 13: Desempenho dos *datasets* usando taxa de erro verdadeira.

Janela	**Dataset(ID)**	p_u		**Amostras**		**%Amostras**		λ		**ROCefficiency**	
		fixa	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*	*fixa*	*móvel*
$\tau = 20$	Lost 1.c4.f	18	-	51983	-	12,1	-	2,67e-18	-	0,623	-
	Lost 17.c4.f	25	-	49688	-	8,9	-	4,47e-15	-	0,483	-
	Ped2.f	34	-	1991	-	0,4	-	1,18e-15	-	0,419	-

Em adição, observando que a métrica de *ROCefficiency* se põe dentro

de uma região considerada boa segundo os critérios de Powers (2011) para classificadores binários, fez-se uso dos resultados alcançados por (MAHADEVAN et al., 2010) para estabelecer comparativos entre os classificadores avaliados por esses autores e o resultante desta simulação. A Figura 46 ilustra o posicionamento da performance alcançada frente ao trabalho realizado por esses autores e outras versões discutidas por eles. Vale destacar que o comparativo é válido se for tomado como base somente a qualidade das inferências relativas ao índice de *informedness*, uma vez que todos os trabalhos comparados estão na linha de abordagem baseada em movimento e não em rastreamento. Esse comparativo mostra que o índice conquistado de *informedness* na presente abordagem, se situa dentro de plausabilidade de uso quando se observam índices similares de trabalhos recentes, embora com abordagens baseadas em movimento e análise baseada em região.

Figura 46: Comparativo de desempenho do vídeo Ped2 com outros trabalhos de DMA baseados em região.



Fonte: Figura original fornecida por Mahadevan et al. (2010) e atualizada com a curva ROC da Figura 45(b).

Outro dado que merece destaque em relação a abordagem apresentada no presente trabalho é o custo computacional exigido entre as soluções. Segundo Mahadevan et al. (2010), usando um computador com desempenho similar ao que se utilizou aqui, eles necessitaram em torno de 2 horas para a fase de treinamento e ainda outros 25 segundos por *frame* na fase de teste. O modelo proposto aqui exigiu pouco mais de 25 minutos em linguagem in-

terpretada do MATLAB® para o treinamento e a busca pelo melhor fator de grade e λ na fase de teste entre 50 valores de p_u. Esta desvantagem do alto custo computacional comuns em modelos de análise baseados em região, conforme discutidas nos capítulo 1 e 2, foi uma da principais motivações pela escolha da abordagem baseada em rastreamento na DMA. Os autores Saligrama e Chen (2012) apresentaram soluções mais recentes e melhoradas na linha de análise baseada em região para vários tipos de *datasets* incluindo o Ped1 da UCSD. Eles contribuíram com estratégias que requerem recursos computacionais menores mas infelizmente o video Ped2 não entrou na avaliação. Um trabalho mais recente como o dos autores Guo et al. (2013), ainda na linha de abordagem baseada em movimento e análise de região, traz melhoras significativas em relação ao trabalho de Mahadevan et al. (2010) tanto em desempenho da DMA quanto no esforço computacional. Embora as evoluções conquistadas por esses autores demonstra incrementos de qualidade nas inferências, ainda falham no quesito de velocidade de resposta na fase de teste. Os vídeos da UCSD que foram amostrados a uma taxa de $\sim$30 FPS, no melhor dos resultados de Guo et al. (2013) é 17 vezes mais rápido que a proposta de (MAHADEVAN et al., 2010) no entanto somente consegue produzir inferências a uma taxa de 0.67FPS. Essa taxa pode ainda ser distante de uma aplicação apropriada em tempo real. Neste ponto, mesmo com um rendimento abaixo dos autores citados a abordagem baseada em rastreamento proposta aqui ganha vantagens no que se refere ao propósito do uso em aplicações de videovigilância real pois as inferências são calculadas em uma complexidade computacional $O(M)$ onde M depende de τ e o número de objetos no *frame*. Na fase de teste as operações de DMA, usando MATLAB®, foram realizas em $\sim$38FPS, ou seja, dentro do intervalo de captação entre um *frame* e outro. Obviamente que essa é somente a performance computacional da etapa de análise de movimento. No entanto, baseado nesse desempenho, é factível dizer que, usando linguagens compiladas na construção dos algorítmos, existe folga para computacionalmente tratar as etapas anteriores de um *framework* baseado em rastreamento. Essa condição leva a garantia da aplicabilidade desta estratégia para o mundo real.

5.2 AMPLIANDO O MÉTODO PARA USO EM MOSAICO DE CÂMERAS

Inspirado na abordagem da DMA proposta no presente trabalho aplicada para câmera única, uma projeção da mesma metodologia e estratégias adotadas podem ser direcionadas para aplicações com múltiplas câmeras típicas de sistemas centralizados e legados de videovigilância. Em uma ideia inicial, a fusão das informações de cada câmera cria agora um vetor (ou matriz) de

regiões formadas pela quantidade de câmeras sincronizadas onde cada objeto móvel vai ocupar qualquer região do grande espaço agora monitorado. A princípio, o uso do modelo de cena com grade fixa passa a conter em cada região da grade, somente os vetores das transições de objetos rastreados entre câmeras. Esses vetores serão gerados por rastreadores aplicados para este propósito ou por anotações em vídeo, similarmente como foi realizado no presente trabalho. Assim, toda a movimentação de objetos entre todas as câmeras participantes de um mosaico de regiões monitoradas, irá formar trajetos definidos aqui como **multilocais**, para a próxima transição ao FOV de outra câmera e **multiglobais** para as próximas τ transições seguintes, de forma similar ao que foi feito para câmera única.

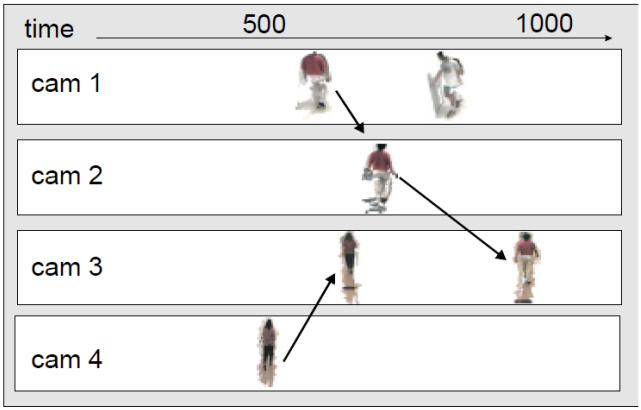
A Figura 47 conforme Kettnaker e Zabih (1999) ilustra o que é um comportamento anormal multilocal. O objeto monitorado na câmera 1 pode possuir comportamento normal local e global na área coberta pela câmera 1. Ao sair do campo de visão da câmera 1 para a câmera 2 o sistema pode identificar uma anormalidade local uma vez que este objeto normalmente deveria circular para o campo de visão da câmera 3 antes de chegar ao campo de visão da câmera 2. Analogamente, caso o movimento de um objeto siga um deslocamento normal entre algumas câmeras e em seguida adote um caminho adverso no campo de visão das câmeras seguintes, ele estará se comportando com uma anomalia de movimento multiglobal.

Embora pareça lógico raciocinar sobre esses comportamentos, existem problemas não triviais para serem resolvidos, como por exemplo: como identificar se o objeto que transitou entre as câmeras é o mesmo objeto? Como tratar as intersecções *overlapping* ou as não intersecções ou oclusões dos campos de visão das câmeras? Alguns caminhos possíveis podem ser inspirados em propostas como a de Javed et al. (2003) ou mesmo usando as redes bayesianas em Kettnaker e Zabih (1999).

A estratégia adotada na DMA proposta permite refletir sobre a aplicação da mesma abordagem quando do uso de outras câmeras em um arranjo organizado. Considerando que as múltiplas câmeras são adotadas em um sistema de monitoração para ampliar a área de visão da cena, pressupõe-se que a captura de suas imagens deve obrigatoriamente estar em sincronismo no ponto de convergência da análise dessas imagens. Os recursos da transmissão desses sinais por redes específicas a partir do ponto de captura como CATV ou IP deve garantir esse alinhamento de tempo.

A independência de contexto encontrada no presente trabalho, também pode levar a outras situações de interesse para monitoração de múltiplos objetos móveis em múltiplas visões como na análise de imagens de microscopia ou mesmo em astronomia. Outra situação adaptável nesta linha são aquelas relacionadas com uma imagem de câmeras megapixel como por exem-

Figura 47: Análise do movimento em múltiplas câmeras.



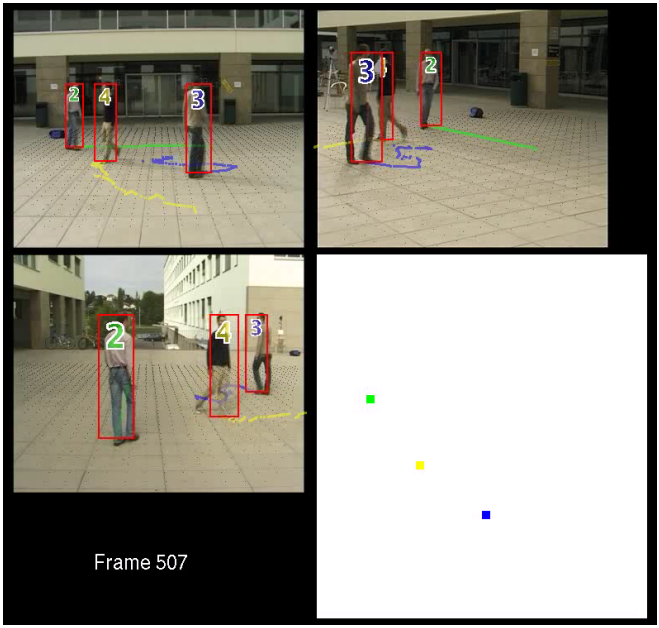
Fonte: (KETTNAKER; ZABIH, 1999).

plo o padrão SXGA (1280x1024 *pixels*) ou WUQSXGA (4200x2690 *pixels*) onde o tratamento de algoritmos em tempo real deve-se tornar dispendioso ou mesmo intratável. Nesse caso pode-se dividir a imagem de alta resolução em múltiplos pedaços com resoluções menores, e tratar cada pedaço como se fosse a captura de imagem de uma câmera exclusiva.

A Figura 48 mostra um outro exemplo do uso de múltiplas câmeras onde múltiplos objetos estão circulando entre elas. Existe uma exigência para que sincronismo entre as câmeras seja estabelecido como um *timestamp* comum entre todas as câmeras. Segundo Berclaz et al. (2008), a robustez no rastreamento de múltiplos objetos é uma das grandes vantagens desta estratégia além de permitir um controle maior sobre o problema da oclusão.

Avaliando o comportamento das transições dos objetos entre as zonas de visão ação de cada câmera observa-se que é possível inferir sobre o anomalias *multilocais* ou *multiglobais* da mesma forma que se faz a avaliação local e global dentro do campo de visão de uma única câmera. O movimento de objetos em múltiplas câmeras pode apresentar regularidades locais e globais de deslocamento em um campo de visão de uma câmera, mas a presença de objetos em áreas cobertas por outras câmeras pode ser encarado como uma anormalidade.

Figura 48: Exemplo de aplicação com múltiplas câmeras.



Fonte: (BERCLAZ et al., 2008)

O capítulo seguinte faz uma análise geral de todas as informações e resultados alcançados no desenvolvimento da presente tese na busca de responder aos objetivos inicialmente traçados do presente trabalho. Tratando-se de um trabalho de cunho empírico-científico, tendências observadas serão o norte para justificar muitos comportamentos e dados mesmo que não tenham sido amplamente estudados ou analiticamente provados. Por se tratar de uma abordagem baseada em ferramental estatístico, incertezas oferecem o tom das discussões.

6 CONCLUSÃO

O presente trabalho apresentou uma nova abordagem para a DMA em cenas do mundo real de videovigilância. Os modelos de cena e de movimento foram planejados e implementados em abordagens baseadas em rastreamento de objetos que transitam entre regiões com agrupamentos uniformes dispostos em grade fixa ou móvel sobre a área da ROI. Adotou-se um modelo estatístico de aprendizagem GMM com treinamento através de um algoritmo EM o qual, através de uma estratégia iterativa, apoiada por um classificador binário baseado em curva ROC, encontra o tamanho de agrupamento das regiões da grade mais adequado para cada cenário avaliado. Todo o modelamento da abordagem foi direcionado na busca de conjuntos de amostras por região que melhor representassem os cenários e o desempenho do reconhecimento de padrões de movimentos sob o menor esforço computacional possível. Para avaliar o desempenho dos modelos utilizados, foram realizadas várias simulações sobre quatro diferentes *datasets* a partir de anotações de vídeo realizadas por ferramental apropriado. Para manter uma relação com trabalhos anteriores ou futuros, todas as simulações utilizaram o conjunto de amostras limitadas as disponíveis nos *datasets* originais, mesmo diante da observação de que em alguns casos, tendências levavam a hipótese de melhora de resultados se amostras adicionais fossem agregadas na análise ou se anotações fieis substituíssem as existentes.

A estratégia montada para a modelagem da abordagem levou em consideração os quesitos para atenuar ou equilibrar os problemas de dimensionalidade, *overtraining* e *overfitting*. Nesse objetivo, a adoção da redução para 3 dimensões do espaço de características dos objetos e a ideia de incrementar iterativamente o fator de grade das regiões uniformes a medida que se mede a qualidade das inferências do modelo, foi fundamental para compreender o comportamento do aprendizado do modelo frente ao resultado de todas as simulações. Embora tenha permitido alcançar algum grau de generalização, a simplicidade da determinação do limiar do classificador binário λ adotado, libertou o DMA das complexidades computacionais intrínsecas do *overfitting*. Da mesma forma, o descarte de amostras que representam o movimento dos objetos dentro dos limites de cada agrupamento da grade, contribuiu de forma positiva para manter em cada região, somente o conjunto de exemplos de padrões de movimento com maiores similaridades intra-classe e menores similaridades inter-classe, acelerando a convergência e precisão do treinamento do GMM. Isso atenuou o problema do *overtraining* mas levou o classificador a se adaptar melhor a classe dominante de amostras do treinamento, contribuindo também para piorar a capacidade de generalização do modelo.

No entanto, mesmo sendo um resultado satisfatório e com desempenho comparável aos trabalhos mais recentes de outros autores em termos do conceito de *informedness*, a capacidade de generalização não foi tratada como objetivo principal do presente trabalho.

Apesar da abordagem do presente trabalho ter sido montada sobre um treinamento baseado em reconhecimento onde o mesmo conjunto de treinamento também participa na fase de teste, os resultados mostraram que o método foi eficaz para vídeos de longa duração e também mostrou comportamento similar para diferentes cenários, contextos, quantidade e qualidade das amostras. Na maioria das simulações o número de amostras iniciais ficou bem abaixo de um milhão das quais somente uma parcela delas foi suficiente para otimizar o desempenho da DMA. Por consequência a abordagem consumiu um esforço computacional reduzido devido à queda exponencial de amostras provocada pelo agrupamento de regiões. A redução experimentada em todas as simulações chegou em limites pouco superiores a $\sim 60\%$ do número inicial de amostras, ou seja, quando $p_u = 1$.

A redução do custo computacional, fundamental para aplicações em tempo real, foram alcançadas devido a três principais premissas assumidas na modelagem da abordagem: *i)* a realização da análise de vídeo a partir de um vetor de 3 dimensões formado por um único descritor do objeto (tipo), sua localização conforme a posição no espaço de regiões e o registro de tempo decorrido desde sua aparição na cena. Esse vetor desvinculou as relações de precisão de formas, cores ou texturas dos objetos e simplificou o modelo multivariado afastando os problemas da "maldição de dimensionalidade" apresentada na seção 2.4; *ii)* a redução do número de amostras devido ao descarte daquelas que fazem parte do rastreamento de um objeto que possui seu centroide se movimentando dentro de um mesmo agrupamento uniforme de *pixels* da ROI da cena. A exclusão dessas amostras mostrou que movimentações menores que as dimensões de largura ou altura dos objetos não afeta a análise do movimento global do objeto e por último *iii)* a limitação da quantidade de objetos móveis monitorados simultaneamente. O número de objetos monitoráveis estabeleceu o limite onde termina a análise de movimento particular e onde começa a análise de movimento denso de objetos (ou de multidão). A quantidade de objetos multiplica o uso de recursos computacionais e cria novos desafios para tratar oclusões, rastreamento e o próprio contexto.

O fato da modelagem da DMA estar dependente somente do deslocamento no espaço-tempo de objetos previamente rastreados e classificados, há uma notória desvinculação do modelo com o contexto da cena monitorada. Essa dissociação motiva enormemente a aplicação dessa estratégia em sistemas legados de videovigilância ou quaisquer outros dependentes somente da análise de trajetória.

A comparação dos resultados obtidos a partir dos dois tipos de distribuição dos agrupamentos uniformes de região em um grade fixa e outra móvel, mostrou que a mobilidade da grade sobre a ROI oferece resultados até 4% melhores em eficiência na DMA como também resulta em limiares de detecção λ mais altos. Diante de que esses valores são resultantes do cálculo de probabilidades, valores mais altos indicam que um número maior de amostras participou de forma mais discriminante na formação de *clusters* do GMM. A grade móvel exige mais esforço computacional e um algoritmo mais elaborado. No entanto, essa complexidade é necessária apenas uma vez para cada ciclo de sua formação. Uma vez calculada, o gabarito de coordenadas de cada região fica disponível até que seja necessária uma nova rodada.

6.1 CONTRIBUIÇÕES DA TESE

A estratégia apresentada seguiu um caminho original diante do estado da arte na DMA. Propostas anteriores não fizeram abordagem com a mesma ideia central de dividir o cenário de vídeo em uma grade de agrupamentos uniformes e ajustados sobre a ROI onde, a partir daí, o processo de análise do movimento se desdobra. As contribuições da tese tornam-se mais evidentes ao adotar-se a metodologia para encontrar o melhor arranjo e tamanho dos agrupamentos de *pixels* que otimiza o desempenho da DMA.

Ao longo do desenvolvimento do presente trabalho, identificou-se uma contribuição central e outras secundárias.

6.1.1 Contribuição Central

A DMA proposta aqui, usou um combinado de estratégias novas no modelamento de cena através da grade móvel; no modelamento da aprendizagem através do uso das propriedades da curva ROC como um classificador binário e no modelamento de movimento através do descarte de amostras redundantes dentro de agrupamentos uniformes ótimos. Os modelos aproveitaram os pontos positivos encontrados na revisão bibliográfica feita até aqui, norteadas pela meta do uso mínimo de informação que possa manter ou mesmo melhorar as inferências da análise de vídeo aplicada especialmente para fins de videovigilância.

6.1.2 Contribuições Secundárias

O uso da curva ROC como sintonizador do limiar de decisão dentre todos os valores de probabilidade calculados a partir de *pdf* multivariadas modeladas por GMM se mostrou um casamento promissor para criar um classificador binário mais confiável.

Observando a relação entre as dimensões dos objetos com o melhor valor de p_u para cada tipo de cenário, é factível inferir que a determinação da faixa ideal de tamanho e arranjo dos agrupamentos ficou fortemente atrelado às dimensões de largura ou altura da maioria dos objetos rastreados. Ou seja, apesar de existir um agrupamento ótimo para cada cenário e para cada tipo de treinamento supervisionado (com taxa de erro aparente ou taxa de erro verdadeira), foi possível estabelecer para abordagens baseadas em rastreamento, uma relação ainda que empírica, de que o tamanho ideal de agrupamento converge para o tamanho médio da dimensões dos objetos rastreados.

Ficou demonstrado nas simulações que a precisão dos resultados da DMA melhorou na comparação entre do modelo de cena com grade fixa, comumente usada em análise de vídeo baseada em região, para o modelo com grade móvel, mesmo usando critérios heurísticos para iniciar o posicionamento de cada agrupamento na ROI.

A metodologia ajudou a diminuir consideravelmente a carga computacional em virtude da redução de dimensionalidade e quantidade dos dados envolvidos em todos os processos de treinamento e teste da DMA.

6.2 TRABALHOS RELACIONADOS DO AUTOR

A essência dos modelos de cena usando grade fixa, de movimento e de aprendizagem da abordagem da DMA discutida aqui, foram resumidas em um artigo pulicado e apresentado no congresso internacional anual ICPRAM2014 (CASAGRANDE; STEMMER, 2014b) sob o foco de um modelo de DMA que consome poucos recursos computacionais. Na sequência, a mesma abordagem, porém utilizando grade móvel foi publicada no congresso internacional ICCCV2014 (CASAGRANDE; STEMMER, 2014a).

6.3 FUTUROS TRABALHOS

A navegação sobre vários temas e o ferramental matemático e computacional associados à análise de vídeo e mais especificamente aqueles voltados para a análise de movimentos anormais revelou muitas possibilidades da

pesquisa nessa área. Mais evidente é a escassez de trabalhos voltados para análise de vídeos de longa duração, especialmente porque esses demandam uso de abordagens baseadas em rastreamento, as quais atualmente ainda possuem muitas questões em aberto.

A divisão da área de análise em agrupamentos orientados por uma grade de regiões produziu ótimos resultados no presente trabalho e dá indicativos de que vale a pena continuar trabalhando com outras formas de segmentar a área da ROI com arranjos que consigam agrupar amostras pelas suas similaridades intra-classes e disjunções inter-classes. Outras formas poligonais de segmentação que são voltadas especialmente para segmentação de imagens podem ser úteis também para o modelamento de cena tais como: Gaussianization proposto por Condurache e Mertins (2013), grades triangulares adaptativas (CONDELL et al., 2002) e o conceito de *superpixel* onde "sementes" são usadas para iniciar um processo de crescimento de região baseado na similaridade de dados da vizinhança. O uso do *dataset* do LOST foi desafiador por se tratar de cenas reais e de longa duração. Uma vez que os vídeos não possuem qualquer separação entre conjuntos de exemplos de treinamento e de teste, eles passam a ser adequados para análise de vídeo não supervisionada. Outra linha de *datasets* voltados para aplicação da pesquisa em videovigilância é apresentada por Oh et al. (2011). Com o mesmo intuito, existem os *datasets* do CVER (Continuous Visual Event Recognition) que incorpora uma série de novidades incluindo maiores resolução de câmeras e número de eventos. Criar cenários reais incluindo múltiplas câmeras para testar a abordagem proposta aqui juntamente com a análise multilocal/multiglobal também é uma sugestão para futuros trabalhos.

Alguns pontos negativos observados em relação a performance e estabilidade de convergência no algoritmo EM e o aprimoramento do modelo GMM e do classificador binário adotados podem ser trabalhados para se conseguir resultados melhores e/ou mais rápidos, especialmente para atender a característica de generalização do DMA.

Seria conveniente encontrar uma explicação ou prova teórica da tendência revelada nos resultados onde os melhores valores de agrupamento dos dois tipos de grade, se situam em uma faixa que se estende até o valor médio das dimensões do *bounding box* dos objetos monitorados. Essas informações não participaram de forma explícita no treinamento, mas emergiram como um tendência que pode ser generalizadas para auxiliar na escolha mais segura em termos de eficiência, nas estratégias de autores que adotam em suas propostas, análise baseada em região.

A estratégia de modelagem descrita nesta tese apenas representa os passos previstos do que deve ser sua implementação no mundo real. Apesar de não ser ser o objetivo desta tese, é possível refletir sobre como generalizar

os resultados conquistados para serem úteis em outras áreas de aplicações. A análise do comportamento do movimento de objetos móveis também é fundamental em sequências de vídeo que trazem respostas importantes para seu fim como: sistemas de produção, automação, análise de tráfego aéreo e urbano, esportes, astronomia, cinemática, mecânica dos fluídos, cinesiologia, agropecuária, zoologia, biologia, entre outros.

REFERÊNCIAS

ABRAMS, A.; TUCEK, J.; LITTLE, J.; JACOBS, N.; PLESS, R. LOST: Longterm Observation of Scenes (with Tracks). In: **Applications of Computer Vision (WACV), 2012 IEEE Workshop on**: WACV, 2012. p. 297–304. ISSN 1550-5790.

ANDERSON, C. J. One look into the future of CMOS chip design. **ISPD**, ACM, p. 1–2, 2009.

APPIAH, K.; HUNTER, A.; OWENS, J.; AIKEN, P.; LEWIS, K. Autonomous real-time surveillance system with distributed IP cameras. In: **Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on**, 2009. p. 1–8.

BANG, J.; KIM, D.; EOM, H. Motion Object and Regional Detection Method Using Block-Based Background Difference Video Frames. In: **Embedded and Real-Time Computing Systems and Applications (RTCSA), 2012 IEEE 18th International Conference on**, 2012. p. 350–357. ISSN 1533-2306.

BARNICH, O.; Van Droogenbroeck, M. ViBe: A Universal Background Subtraction Algorithm for Video Sequences. **Image Processing, IEEE Transactions on**, v. 20, n. 6, p. 1709–1724, June 2011. ISSN 1057-7149.

BASHARAT, A.; GRITAI, A.; SHAH, M. Learning object motion patterns for anomaly detection and improved object detection. In: **Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on**, 2008. p. 1–8. ISSN 1063-6919.

BERCLAZ, J.; FLEURET, F.; FUA, P. Multi-camera Tracking and Atypical Motion Detection with Behavioral Maps. In: **ECCV (3)**, 2008. p. 112–125.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006. ISBN 978-0-387-31073-2.

CANNONS, K. **A Review of Visual Tracking**. Technical report. Toronto, Ontario Canada, set. 2008. CSE-2008-07, 235 p.

CASAGRANDE, J. H. B.; STEMMER, M. R. Abnormal Motion Analysis for Tracking-Based Approaches Using Region-Based Method with Mobile Grid. **Abnormal Motion Analysis for Tracking-Based Approaches Using Region-Based Method with Mobile Grid**, Journal of Image and Graphics, v. 2, n. 1, p. 22–27, jun. 2014.

CASAGRANDE, J. H. B.; STEMMER, M. R. Region-Based Abnormal Motion Detection in Video Surveillance. In: **ICPRAM**. Presented at ICPRAM Angers, France, 6-8 Mar, 2014: SciTePress, 2014.

CHEN, W.-T.; CHEN, P.-Y.; LEE, W.-S.; HUANG, C.-F. Design and Implementation of a Real Time Video Surveillance System with Wireless Sensor Networks. In: **VTC Spring**: IEEE, 2008. p. 218–222.

CONDELL, J.; SCOTNEY, B.; MORROW, P. Detection and estimation of motion using adaptive grids. In: **Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on**, 2002. v. 2, p. 675–678 vol.2.

CONDURACHE, A. P.; MERTINS, A. Accelerated Nonlinear Gaussianization for Feature Extraction. In: MARSICO, M. D.; FRED, A. L. N. (Ed.). **ICPRAM**: SciTePress, 2013. p. 121–126. ISBN 978-989-8565-41-9.

CONG, Y.; YUAN, J.; TANG, Y. Video Anomaly Search in Crowded Scenes via Spatio-Temporal Motion Context. **IEEE Transactions on Information Forensics and Security**, v. 8, n. 10, p. 1590–1599, 2013.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An Introdution to Support Vector Machines and Other Kernel-Based Learning Methods**: Cambridge University Press, 2000. ISBN 0-521-78019-5.

CZYZEWSKI, A.; DALKA, P. Moving Object Detection and Tracking for the Purpose of Multimodal Surveillance System in Urban Areas. In: TSIHRINTZIS, G. A.; VIRVOU, M.; HOWLETT, R. J.; JAIN, L. C. (Ed.). **New Directions in Intelligent Interactive Multimedia**: Springer, 2008, (Studies in Computational Intelligence, v. 142). p. 75–84. ISBN 978-3-540-68126-7.

ELHOSEINY, M.; BAKRY, A.; ELGAMMAL, A. MultiClass Object Classication in Video Surveillance Systems Experimental Study. **IEEE Conference on Computer Vision and Pattern Recognition Workshops**, CVPR'13, p. 788–793, 2013.

ERMIS, E. B.; SALIGRAMA, V.; JODOIN, P.-M.; KONRAD, J. Motion Segmentation and Abnormal Behavior Detection via Behavior Clustering. In: **ICIP**: IEEE, 2008. p. 769–772.

EZZAHOUT, A.; THAMI, R. Conception and development of a video surveillance system for detecting, tracking and profile analysis of a person. In: **ISKO-Maghreb, 2013 3rd International Symposium**, 2013. p. 1–5.

FANG, L.; MENG, Z.; CHEN, C.; HUI, Q. Smart Motion Detection Surveillance System. In: **Education Technology and Computer, 2009. ICETC '09. International Conference on**, 2009. p. 171–175.

FEIZI, A.; AGHAGOLZADEH, A.; SEYEDARABI, H. Behavior recognition and anomaly behavior detection using clustering. In: **Telecommunications (IST), 2012 Sixth International Symposium on**, 2012. p. 892–896.

FIGUEIREDO, M. A. T.; JAIN, A. Unsupervised learning of finite mixture models. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, v. 24, n. 3, p. 381–396, 2002. ISSN 0162-8828.

GONG, S.; LOY, C. C.; XIANG, T. Security and Surveillance. In: MOESLUND, T. B.; HILTON, A.; KRÜGER, V.; SIGAL, L. (Ed.). **Visual Analysis of Humans**: Springer, 2011. p. 455–472. ISBN 978-0-85729-996-3.

GONZALEZ, R. C.; WOODS, R. E. **Digital image processing**. 3rd. ed. Upper Saddle River, N.J.: Prentice Hall, 2008. ISBN 9780131687288 013168728X 9780135052679 013505267X.

GRYN, J.; WILDES, R.; TSOTSOS, J. Detecting Motion Patterns via Direction Maps with Application to Surveillance. In: **Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on**, 2005. v. 1, p. 202–209.

GUO, Z.; LI, N.; XU, D.; CHEN, Y.-L.; WU, X.; GAO, Z. A novel statistical learning-based framework for automatic anomaly detection and localization in crowds. In: **Robotics and Biomimetics (ROBIO), 2013 IEEE International Conference on**, 2013. p. 1211–1215.

HAMPAPUR, A.; BORGER, S.; BROWN, L.; CARLSON, C.; CONNELL, J.; LU, M.; SENIOR, A.; REDDY, V.; SHU, C.; TIAN, Y. S3: The IBM Smart Surveillance System: From Transactional Systems to Observational Systems. In: **Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on**, 2007. v. 4, p. IV–1385–IV–1388. ISSN 1520-6149.

HANAPIAH, F.; AL-OBAIDI, A.; CHAN, C. S. Anomalous trajectory detection using the fusion of fuzzy rule and local regression analysis. In: **Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on**: ISSPA, 2010. p. 165–168.

HAQUE, M.; MURSHED, M. Abnormal Event Detection in Unseen Scenarios. In: **Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on**, 2012. p. 378–383.

HARRIS, C.; STEPHENS, M. A combined corner and edge detector. In: **In Proc. of Fourth Alvey Vision Conference**, 1988. p. 147–151.

HU, W.; TAN, T.; WANG, L.; MAYBANK, S. J. A survey on visual surveillance of object motion and behaviors. **IEEE Transactions on Systems, Man, and Cybernetics, Part C**, v. 34, n. 3, p. 334–352, 2004.

HU, W.; XIAO, X.; FU, Z.; XIE, D.; TAN, T.; MAYBANK, S. J. A System for Learning Statistical Motion Patterns. **IEEE Trans. Pattern Anal. Mach. Intell.**, v. 28, n. 9, p. 1450–1464, 2006.

HU, X.; TANG, Y.; ZHANG, Z. Video object matching based on SIFT algorithm. In: **Neural Networks and Signal Processing, 2008 International Conference on**, 2008. p. 412–415.

HUANG, K.; WANG, S.; TAN, T.; MAYBANK, S. Human Behavior Analysis Based on a New Motion Descriptor. **Circuits and Systems for Video Technology, IEEE Transactions on**, v. 19, n. 12, p. 1830–1840, Dec 2009. ISSN 1051-8215.

Jacques Junior, J.; Raupp Musse, S.; JUNG, C. Crowd Analysis Using Computer Vision Techniques. **Signal Processing Magazine, IEEE**, v. 27, n. 5, p. 66–77, Sept 2010. ISSN 1053-5888.

JAVED, O.; RASHEED, Z.; ALATAS, O.; SHAH, M. KNIGHT: a real time surveillance system for multiple and non-overlapping cameras. In: **ICME**: IEEE, 2003. p. 649–652. ISBN 0-7803-7965-9.

JAVED, O.; SHAH, M. **Automated Multi-Camera Surveillance: Algorithms and Practice.**: Springer, 2008. (The International Series in Video Computing, v. 10). ISBN 978-0-387-78881-4.

JIANG, F.; YUAN, J.; TSAFTARIS, S. A.; KATSAGGELOS, A. K. Anomalous video event detection using spatiotemporal context. **Computer Vision and Image Understanding**, v. 115, n. 3, p. 323–333, 2011.

KAVASIDIS, I.; PALAZZO, S.; SALVO, R. D.; GIORDANO, D.; SPAMPINATO, C. An Innovative Web-based Collaborative Platform for Video Annotation. **Multimedia Tools Appl.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 70, n. 1, p. 413–432, maio 2014. ISSN 1380-7501.

KETTNAKER, V.; ZABIH, R. Bayesian multi-camera surveillance. In: **Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.**, 1999. v. 2, p. –259 Vol. 2. ISSN 1063-6919.

KIRYATI, N.; RAVIV, T.; IVANCHENKO, Y.; ROCHEL, S. Real-time abnormal motion detection in surveillance video. In: **Pattern Recognition, 2008. ICPR 2008. 19th International Conference on**, 2008. p. 1–4. ISSN 1051-4651.

KO, T. A survey on behavior analysis in video surveillance for homeland security applications. In: **AIPR**: IEEE Computer Society, 2008. p. 1–8. ISBN 978-1-4244-3125-0.

KWON, E.; NOH, S.; JEON, M.; SHIM, D. Scene Modeling-Based Anomaly Detection for Intelligent Transport System. In: **Intelligent Systems Modelling Simulation (ISMS), 2013 4th International Conference on**, 2013. p. 252–257. ISSN 2166-0662.

LAVEE, G.; RIVLIN, E.; RUDZSKY, M. Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video. **IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS**, v. 39, n. 5, p. 489–504, set. 2009.

LI, B.; TIAN, B.; LI, Y.; XIONG, G. Design and implementation of the networked video surveillance and management platform in Suzhou subway line 1. In: **Service Operations and Logistics, and Informatics (SOLI), 2013 IEEE International Conference on**, 2013. p. 136–141.

LI, H.; ACHIM, A.; BULL, D. Unsupervised video anomaly detection using feature clustering. **Signal Processing, IET**, v. 6, n. 5, p. 521–533, 2012. ISSN 1751-9675.

LI, J.; GONG, S.; XIANG, T. Learning Behavioural Context. **International Journal of Computer Vision**, v. 97, n. 3, p. 276–304, 2012.

LI, R.; ZHU, L.; YU, S.-S. Intelligent Video Monitor System Based on Neural Networks Analysis. In: **Wireless Communications, Networking and Mobile Computing, 2009. WiCom '09. 5th International Conference on**, 2009. p. 1–6.

LI, Y.; YIN, Y. Towards Suspicious Behavior Discovery in Video Surveillance System. In: **Knowledge Discovery and Data Mining, 2009. WKDD 2009. Second International Workshop on**, 2009. p. 539–541.

LIU, C.; WANG, G.; NING, W.; LIN, X.; LI, L.; LIU, Z. Anomaly detection in surveillance video using motion direction statistics. In: **ICIP**: IEEE, 2010. p. 717–720. ISBN 978-1-4244-7994-8.

MA, Z.; WAN, J. Survey of Data Association of Moving Objects tracking in Video Sensors network. **The Ninth International Conference on Electronic Measurement & Instruments**, p. 250–254, 2009.

MAGGIO, E.; CAVALLARO, A. Learning Scene Context for Multiple Object Tracking. **Image Processing, IEEE Transactions on**, v. 18, n. 8, p. 1873–1884, Aug 2009. ISSN 1057-7149.

MAHADEVAN, V.; LI, W.; BHALODIA, V.; VASCONCELOS, N. Anomaly detection in crowded scenes. In: **Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on**, 2010. p. 1975–1981. ISSN 1063-6919.

MOREELS, P.; PERONA, P. Evaluation of features detectors and descriptors based on 3D objects. In: **Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on**, 2005. v. 1, p. 800–807 Vol. 1. ISSN 1550-5499.

MORRIS, B.; TRIVEDI, M. A Survey of Vision-Based Trajectory Learning and Analysis for Surveillance. **Circuits and Systems for Video Technology, IEEE Transactions on**, v. 18, n. 8, p. 1114–1127, Aug 2008. ISSN 1051-8215.

NARAYANA, M.; HAVERKAMP, D. A Bayesian algorithm for tracking multiple moving objects in outdoor surveillance video. In: **Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on**, 2007. p. 1–8. ISSN 1063-6919.

NAZARE, A.; SANTOS, C. dos; FERREIRA, R.; Robson Schwartz, W. Smart surveillance framework: A versatile tool for video analysis. In: **Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on**, 2014. p. 753–760.

OH, S.; HOOGS, A.; PERERA, A.; CUNTOOR, N.; CHEN, C.-C.; LEE, J. T.; MUKHERJEE, S.; AGGARWAL, J.; LEE, H.; DAVIS, L.; SWEARS, E.; WANG, X.; JI, Q.; REDDY, K.; SHAH, M.; VONDRICK, C.; PIRSIAVASH, H.; RAMANAN, D.; YUEN, J.; TORRALBA, A.; SONG, B.; FONG, A.; ROY-CHOWDHURY, A.; DESAI, M. A large-scale benchmark dataset for event recognition in surveillance video. In: **Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on**, 2011. p. 3153–3160. ISSN 1063-6919.

PANDA, D.; MEHER, S. A Gaussian mixture model with Gaussian weight learning rate and foreground detection using neighbourhood correlation. In: **Microelectronics and Electronics (PrimeAsia), 2013 IEEE Asia Pacific Conference on Postgraduate Research in**, 2013. p. 158–163.

PATRICK, R.; BOURBAKIS, N. Surveillance Systems for Smart Homes: A Comparative Survey. In: **Tools with Artificial Intelligence, 2009. ICTAI '09. 21st International Conference on**, 2009. p. 248–252. ISSN 1082-3409.

PONTIL, M.; VERRI, A. Support Vector Machines for 3D Object Recognition. **IEEE Trans. Pattern Anal. Mach. Intell.**, v. 20, n. 6, p. 637–646, 1998.

POPPE, R. A survey on vision-based human action recognition. **Image and Vision Computing**, v. 28, n. 6, p. 976–990, 2010. ISSN 0262-8856.

POWERS, D. M. W. Evaluation: From Precision, Recall and F-Factor to ROC: Informedness, Markedness & Correlation. **Journal of Machine Learning Technologies**, Australia, v. 2, n. Issue 1, p. 37–63, 2011. ISSN ISSN: 2229-3981 & ISSN: 2229-399X,.

RÄTY, T. Survey on Contemporary Remote Surveillance Systems for Public Safety. **Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on**, v. 40, n. 5, p. 493–515, 2010. ISSN 1094-6977.

REVATHI, A. R.; KUMAR, D. A Survey Of Activity Recognition And Understanding The Behavior In Video Survelliance. **CoRR**, abs/1207.6774, 2012.

ROBERTSON, N. M.; REID, I. D. A general method for human activity recognition in video. **Computer Vision and Image Understanding**, v. 104, n. 2-3, p. 232–248, 2006.

RUSSEL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach** . 3rd. ed.: Pearson, 2009. ISBN 978-0136042594.

SALIGRAMA, V.; CHEN, Z. Video anomaly detection based on local statistical aggregates. In: **Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on**, 2012. p. 2112–2119. ISSN 1063-6919.

SHI, J.; TOMASI, C. Good features to track. In: **Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on**, 1994. p. 593–600. ISSN 1063-6919.

SHI, Y.; GAO, Y.; WANG, R. Real-Time Abnormal Event Detection in Complicated Scenes. In: **ICPR**: IEEE, 2010. p. 3653–3656.

SODEMANN, A. A.; ROSS, M. P.; BORGHETTI, B. J. A Review of Anomaly Detection in Automated Surveillance. **IEEE Transactions on Systems, Man, and Cybernetics, Part C**, v. 42, n. 6, p. 1257–1272, 2012.

SUDO, K.; OSAWA, T.; TANAKA, H.; KOIKE, H.; ARAKAWA, K. Online anomal movement detection based on unsupervised incremental learning. In: **Pattern Recognition, 2008. ICPR 2008. 19th International Conference on**, 2008. p. 1–4. ISSN 1051-4651.

SUN, W.; GUO, B.-L. A Robust Object Detecting and Tracking Method. In: **Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on**, 2008. v. 4, p. 121–125.

TEHRANI, M.; KLEIHORST, R.; MEIJER, P.; SPAANENBURG, L. Abnormal motion detection in a real-time smart camera system. In: **Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on**, 2009. p. 1–7.

TITTA, S. de; GERA, G.; MARCENARO, L. VTrack: Video analytics for automatic video-surveillance. In: **AVSS**: IEEE Computer Society, 2011. p. 536–538. ISBN 978-1-4577-0845-9.

TZIAKOS, I.; CAVALLARO, A.; XU, L.-Q. Local Abnormality Detection in Video Using Subspace Learning. In: **Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on**, 2010. p. 519–525.

University of Maryland. **ViPER: The Video Performance Evaluation Resource**. Laboratory for Language and Media Processing - Institute for Advanced Computer Studies, 2005. Disponível em: http://viper-toolkit.sourceforge.net. Acesso em: 20 set. 2014., 2005.

VONDRICK, C.; PATTERSON, D.; RAMANAN, D. Efficiently Scaling up Crowdsourced Video Annotation - A Set of Best Practices for High Quality, Economical Video Labeling. **International Journal of Computer Vision**, v. 101, n. 1, p. 184–204, 2013.

XIANG, T.; GONG, S. Video Behaviour Profiling and Abnormality Detection without Manual Labelling. In: **ICCV**: IEEE Computer Society, 2005. p. 1238–1245. ISBN 0-7695-2334-X.

XIANG, T.; GONG, S. Incremental and adaptive abnormal behaviour detection. **Computer Vision and Image Understanding**, v. 111, p. 59–73, jan. 2008.

XU, X.; TANG, J.; LIU, X.; ZHANG, X. Human behavior understanding for video surveillance: Recent advance. In: **Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on**, 2010. p. 3867–3873. ISSN 1062-922X.

YANG, M.-J.; THAM, J. Y.; WU, D.; GOH, K. H. Cost effective IP camera for video surveillance. In: **Industrial Electronics and Applications, 2009. ICIEA 2009. 4th IEEE Conference on**, 2009. p. 2432–2435.

YE, Y.; CI, S.; KATSAGGELOS, A.; LIU, Y.; QIAN, Y. Wireless Video Surveillance: A Survey. **Access, IEEE**, v. 1, p. 646–660, 2013. ISSN 2169-3536.

YU, S.-Z. Hidden semi-Markov models. **Artif. Intell.**, v. 174, n. 2, p. 215–243, 2010.

YU, T.-H.; MOON, Y.-S. Unsupervised Abnormal Behavior Detection for Real-time Surveillance Using Observed History. **IAPR Conference on Machine Vision Applications**, p. 166–169, 2009.

YUEN, J.; RUSSELL, B.; LIU, C.; TORRALBA, A. LabelMe video: Building a video database with human annotations. In: **Computer Vision, 2009 IEEE 12th International Conference on**, 2009. p. 1451–1458. ISSN 1550-5499.

ZAHARESCU, A.; WILDES, R. Anomalous Behaviour Detection Using Spatiotemporal Oriented Energies, Subset Inclusion Histogram Comparison and Event-Driven Processing. In: DANIILIDIS, K.; MARAGOS, P.; PARAGIOS, N. (Ed.). **ECCV (1)**: Springer, 2010. (Lecture Notes in Computer Science, v. 6311), p. 563–576. ISBN 978-3-642-15548-2.

ZENG, S.; CHEN, Y. Online-learned classifiers for robust multitarget tracking. In: **Neural Networks (IJCNN), The 2011 International Joint Conference on**, 2011. p. 1275–1280. ISSN 2161-4393.

ZHANG, S.; CHAN, S.; QIU, R. D.; NG, K.; HUNG, Y.; LU, W. On the design and implementation of a high definition multi-view intelligent video surveillance system. In: **Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on**, 2012. p. 353–357.

ZHANG, T.; WILIEM, A.; LOVELL, B. C. Region-Based Anomaly Localisation in Crowded Scenes via Trajectory Analysis and Path Prediction. In: **DICTA**: IEEE, 2013. p. 1–7.

ZHANG, Y.; LIU, K.; YANG, J. General moving objects recognition method based on graph embedding dimension reduction algorithm. **Journal of Zhejiang University SCIENCE A**, v. 10, n. 7, p. 976–984, 2009.