Caio de Jesus Gregoratto

Detecção de Comportamento Anormal em Vídeos de Multidão

Manaus

Maio de 2016

Caio de Jesus Gregoratto

Detecção de Comportamento Anormal em Vídeos de Multidão

Trabalho apresentado ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para obtenção do grau de Mestre em Informática.

Univesidade Federal do Amazonas

Instituto de Computação

Programa de Pós-graduação em Informática

Orientadora: Eulanda Miranda dos Santos

Manaus

Maio de 2016

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

G819d

Gregoratto, Caio de Jesus
Detecção de Comportamento Anormal em Vídeos de Multidão / Caio de Jesus Gregoratto. 2016
82 f.: il. color; 31 cm.

Orientadora: Eulanda Miranda dos santos
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Sistemas de Segurança Baseados em Vídeo. 2. Reconhecimento de Comportamento de Multidões. 3. Métodos Baseados na Aparência. 4. Métodos Baseados em Características. I. santos, Eulanda Miranda dos II. Universidade Federal do Amazonas III. Título

Agradecimentos

Gostaria de agradecer aos meus pais Maria de Fátima e Gevaldir Gregoratto e ao meu irmão Rafael de Jesus, por todo o apoio e carinho incondicionais.

Obrigado aos companheiros de estudo e de pesquisa: Michel Yvano, Bernardo Gatto, Adria Menezes, Rayol Neto, Marília Feitoza e tantos outros que me ajudaram a superar os desafios encontrados e fizeram do mestrado uma jornada ainda mais gratificante. São todos amigos que vieram como um verdadeiro presente na minha vida.

Agradeço aos professores Waldir Sabino, José Pio e Eduardo Souto, que sempre estiveram prontos para me auxiliar e direcionar nos momentos de incerteza, ensinando com muito respeito e atenção.

Obrigado aos professores Acauan Ribeiro, Felipe Lobo, Fabio Parreira e Luciano Ferreira que tornaram possível meu ingresso nesse apaixonante desafio que foi o mestrado. Agradeço também por terem depositado em mim uma generosa carga de confiança ao me orientar e motivar para mais esse passo em minha formação.

Gostaria de agradecer à minha orientadora, a professora Eulanda Miranda dos Santos, por me receber sempre com muita paciência e de forma respeitosa em qualquer que fosse a situação. Seus conselhos foram fundamentais para o desenvolvimento deste trabalho e para meu amadurecimento pessoal e acadêmico.

Agradeço ao suporte dado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) e a todos que ajudam a manter o Instituto de Computação (IComp), sempre com um trabalho sério e pontual.

Por fim, agradeço a Deus, por tudo.

Resumo

Sistemas de segurança produzem uma quantidade massiva de material de vídeo que pode ser utilizada para reconhecer comportamento anormal ou atividades que ofereçam riscos à saúde das pessoas. Entretanto, nem sempre os operadores humanos são capazes de avaliar de forma coerente todo o material disponível. Portanto, reconhecer comportamento em vídeo de forma automatizada pode ser fundamental para que o uso de sistemas de vigilância seja eficiente em manter a segurança de uma área ou a integridade das pessoas em locais públicos. Diante disso, esta dissertação apresenta um método voltada para detectar e reconhecer comportamento anormal em vídeos de multidão. Esse método combina técnicas baseadas em características com técnicas baseadas na aparência e as utiliza conforme o contexto das atividades presentes na cena. Técnicas baseadas na aparência utilizam modelos matemáticos gerados a partir dos níveis de intensidade da imagem para realizar suas tarefas, enquanto que as técnicas baseadas em características usam dados extraídos da imagem, como bordas, linhas e coordenadas, para derivar seus modelos. O método proposto exibe para o operador humano, por meio de marcações visuais, somente conteúdo com possíveis ocorrências de aglomeração ou dispersão da multidão, comportamentos considerados anormais avaliados nesta pesquisa. Os resultados obtidos nos experimentos mostram que a abordagem proposta é capaz de reconhecer comportamentos anormais em vídeos de multidão e marcar as regiões na imagem onde ocorrem anomalias do tipo aglomeração ou dispersão das pessoas na cena. O método proposto, diferente das demais abordagens existentes na literatura, faz avaliações distintas entre as cenas suspeitas de conter comportamento anormal e as cenas com comportamento normal ou somente com a imagem de fundo. Como consequência, os resultados dos experimentos mostram que o método proposto apresenta tempo de execução 64% menor do que os *baselines* em uma base de dados criada neste trabalho e 71% menor nas bases de dados UMN e PETS2009. Além disso, o método proposto atinge uma acurácia de 90% na base de dados YAB, enquanto o *baseline* atinge 85%.

Palavras-chave: Sistemas de Segurança Baseados em Vídeo, Reconhecimento de Comportamento de Multidões, Métodos Baseados na Aparência, Métodos Baseados em Características.

Abstract

Security systems produce a massive amount of video material that can be used to recognize abnormal behavior or activities, which expose the people to life-threatening scenarios. However, human operators are not able to evaluate all the material available in a consistent manner. So, the automatic recognition of video behavior can be crucial for the effective use of surveillance systems to maintain the security of an area or the integrity of people in public places. This work presents an method focused on the recognition of abnormal behavior in crowd videos. This method combines feature-based methods with appearance-based methods and use them according to the context of the scene. Appearance-based methods create models based on the levels of image intensity, while feature-based methods use data extracted from the image, such as edges, lines and coordinates, to generate their models. The feature-based approach is generally used because it describes the scene in more details, however it involves higher computational costs. The proposed method displays for the human operator only content with possible crowd agglomeration or dispersing, which are considered abnormal behaviors evaluated in this research. The displayed video receive visual marks to help the human operator to locate suspicious activities identified by the system. The results obtained during the experiments show that the proposed method is able to recognize abnormal behaviors in crowd videos and mark areas of the image where abnormalities as agglomeration or dispersion are detected. The proposed method, different from classical approaches available in the literature, makes assessments of the suspect scenes different from the assessment of the scenes with normal behavior or with only the background. As a consequence, experimental results show that the proposed method performs 64% faster than the baselines over a database created for this work, as well as 71% faster than baselines on UMN and PETS2009 databases. In addition, the proposed method achieves 90% of accuracy on the YAB database, otherwise the baseline method achieves 85% of accuracy.

Keywords: Video-Based Surveillance Systems, Crowd Behavior Recognition, Appearance-based Methods, Feature-based Methods.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

$2D^2$-PCA	*Two-directional two-dimensional Principal Component Analysis*
2D-PCA	*Bi-directional Principal Component Analysis*
CMI	*Crowd Motion Intensity*
DDM	*Drift Detection Method*
EDDM	*Early Drift Detection Method*
EMC	*Embedded Markov Chain*
FO	Fluxo Óptico
GMM	*Gaussian Mixture Model*
HDP	*Hierarchical Dirichlet Process*
HMM	*Hidden Markov Model*
IMED	*Image Euclidean Distance*
IPCA	*Incremental Principal Component Analysis*
LDOF	*Large Displacement Optical Flow*
MSM	*Mutual Subspace Method*
PCA	*Principal Component Analysis*
PE	*Particle Entropy*
PETS2009	PETS *Crowd Sensing Dataset Challenge*
PLW	*Point-light Walker*
PR	*Pattern Recognition*
SSIM	*Structural Similarity Index*
SVBV	Sistema de Vigilância Baseado em Vídeo
SVD	*Singular Value Decomposition*
SVM	*Support Vector Machine*
UMN	*Unusual Crowd Activity Dataset*
YAB	*YouTube Abnormal Behavior*

Sumário

1 Introdução

No intuito de manter a segurança em locais públicos, Sistemas de Vigilância Baseados em Vídeo (SVBV) são instalados em praças, parques, campus de universidades, shoppings entre outros (ZIN et al., 2014). O uso desses sistemas pode prevenir o surgimento de brigas, depredações, roubos ou atos de terrorismo. Para tanto, é necessário que um operador humano responsável pelo sistema de vigilância, em meio a várias telas do circuito de câmeras, identifique o comportamento em tempo hábil para que seja dado início à tomada de decisão adequada (SINDHUJA; SRINIVASAGAN; KALAISELVI, 2014). No entanto, mesmo pessoas treinadas podem não identificar uma determinada atividade suspeita, seja pelo alto nível de fadiga ou pela grande quantidade de imagens avaliadas ao mesmo tempo. O tempo de resposta do operador humano pode variar, seja quando um comportamento é iniciado de forma gradual ou apresentar movimentos bruscos (COHEN et al., 2008; VISHWAKARMA; AGRAWAL, 2012).

Assim, o uso de ferramentas que auxiliem o operador humano na identificação de comportamento suspeito pode reduzir o tempo de resposta diante de um possível cenário de risco (HUEBER et al., 2014). Esse recurso, agregado ao SVBV, pode auxiliar no controle de multidões ou na identificação de objetos abandonados que, em caso de um ato terrorista, por exemplo, podem ser bombas deixadas em meio à multidão (ZIN et al., 2011; ZIN et al., 2014). Independentemente da aplicação, em todos esses exemplos, o sistema deve apresentar uma resposta em tempo real para prevenir ou tratar o cenário de risco. O termo tempo real, neste contexto, refere-se ao intervalo de tempo entre cada quadro do vídeo, que pode variar entre 33 e 100 milissegundos, de acordo com a taxa de quadros por segundo do vídeo ou da câmera (BRONTE et al., 2014).

Sistemas de vigilância capazes de gerar informações adicionais sobre a cena ao operador humano são chamados de sistemas inteligentes. Esses sistemas são chamados inteligentes por implementarem algoritmos de aprendizagem de máquina e técnicas de visão computacional capazes de reconhecer pessoas em uma cena, objetos específicos como malas ou armas, e ações suspeitas como um assalto ou uma briga (GUODONG, 2011; ROSHTKHARI; LEVINE, 2013; RODRÍGUEZ et al., 2014). A atividade ou o objeto reconhecido varia de acordo com o propósito do sistema desenvolvido e com o tipo de problema tratado.

Os sistemas inteligentes são desenvolvidos para auxiliar o operador humano a identificar um comportamento suspeito (COHEN et al., 2008). Nesse contexto, este trabalho propõe uma arquitetura modular para sistemas inteligentes que visa reduzir o custo computacional de SVBVs, especialmente quando estes recebem conteúdo que não seja comportamento suspeito. Dessa forma, a arquitetura proposta potencializa a viabilidade de

implantação de sistemas inteligentes em SVBVs reais. A redução no custo é alcançada ao dividir o sistema em duas fases. A primeira fase, de menor custo, é baseada na aparência, ou seja, avalia os níveis de intensidade dos pixel da imagem. Essa fase fica responsável por avaliar a cena sem qualquer pré-processamento. Essa etapa decide se a segunda fase entrará em atividade ou não. A segunda fase, baseada em características, é responsável por validar o sinal de ativação da primeira fase e, caso confirmada a existência de um comportamento suspeito, exibir para o operador humano em quais áreas da imagem ocorre uma possível anormalidade. O comportamento anormal que este trabalho investiga é de aglomeração ou dispersão de multidões. Assim, as bases de dados utilizadas são compostas por vídeos de sistemas de vigilância com cenas de multidão. A próxima seção descreve mais detalhes do problema tratado neste trabalho e a Seção 1.2 apresenta a justifica para o desenvolvimento desta pesquisa.

1.1 Definição do Problema

Sistemas de Vigilância são geralmente baseados em vídeo e, mesmo em cenários com dimensões pequenas como corredores ou portarias, várias câmeras compõem um único circuito do SVBV. Dessa forma, um conjunto de vídeos deve ser avaliado pelo operador humano na busca por comportamento suspeito (SAINI et al., 2012; BERTINI; BIMBO; SEIDENARI, 2012). O principal fator para justificar o uso de múltiplas câmeras para o mesmo cenário é a tentativa de cobrir todo um ambiente de forma que a oclusão de objetos seja reduzida (SAINI; ATREY; SADDIK, 2014). Com isso, um SVBV com grande número de câmeras gera elevada quantidade de informação de vídeo. Esse conteúdo de vídeo adquirido contém, em sua maioria, cenas apenas com a imagem de fundo ou somente comportamento normal, as quais não precisam ser avaliadas pelo operador humano (POPOOLA; WANG, 2012).

Com o intuito de auxiliar o operador humano, podem ser exibidos somente vídeos que contenham conteúdo suspeito quanto à quebra de segurança. Assim, o operador passaria de dezenas de vídeos para um ou dois vídeos com conteúdo adquirido pela câmera acompanhado de marcações que o auxiliem a identificar a região na cena com comportamento suspeito (BERTINI; BIMBO; SEIDENARI, 2012). Para tanto, a ferramenta responsável por avaliar todos os vídeos deve ter baixo custo de processamento e memória. Neste trabalho, custo será referente ao tempo necessário para processar as informações entre cada quadro recebido do vídeo. O baixo custo da ferramenta que analisa os vídeos em busca de comportamentos anômalos possibilita que sejam mantidas duas características comuns, e até mesmo essenciais para os SVBVs: aquisição das imagens feita por vários vídeos ao mesmo tempo e reconhecimento de comportamento anômalo no menor espaço de tempo possível (COHEN et al., 2008; VISHWAKARMA; AGRAWAL, 2012; LIU; LI; JIA, 2014).

Para prover a primeira característica, um SVBV inteligente pode analisar vários vídeos de maneira independente e destacar aqueles que apresentem comportamento suspeito (COHEN et al., 2008). A segunda característica trata do alerta dado pelo sistema, caso a vigilância seja feita com o apoio de um sistema inteligente, este deve enviar o alerta de comportamento suspeito ao operador humano com o menor atraso possível, para que uma resposta seja dada diante do provável cenário de risco (GUODONG, 2011; VISHWAKARMA; AGRAWAL, 2012).

1.2 Justificativa

Como na maior parte do tempo, vídeos de segurança contêm imagens de fundo ou cenas de comportamento normal (JIANG et al., 2011), esses trechos não necessitam de marcações. Nesses casos, os algoritmos de reconhecimento de comportamento tentam extrair informações de todas as imagens, inclusive da imagem de fundo. Como resultado, o custo da execução do algoritmo em cenários com comportamento normal será o mesmo de quando houver algum comportamento suspeito na cena. Existem ainda propostas que apresentam crescimento no custo de computação conforme o número de pessoas na cena aumenta (SAINI et al., 2012). Assim, como esta pesquisa trata vídeos com multidão, a abordagem baseada em características não avalia o número de pessoas na cena.

Uma forma de reduzir o custo do método é adotar uma arquitetura que comporte a execução de elementos com maior custo apenas em momentos específicos, como cenas suspeitas de conter comportamento anormal. Assim, módulos são executados somente em cenas que contenham elementos referentes à sua função. Selecionar módulos com menor custo para serem executados enquanto não existe comportamento anormal pode resultar em menor custo geral na execução do sistema inteligente. Esta redução do custo é referente ao tempo em que os demais módulos permanecerão inativos, visto que grande parte do conteúdo adquirido pelo SVBV é de comportamento normal (COHEN et al., 2008; JIANG et al., 2011).

Além disso, uma estratégia modular possibilita que sejam combinados métodos com abordagens distintas, e extrair o que cada tipo de método tem a oferecer de melhor. Por exemplo, se os vídeos exibem comportamento normal ou simplesmente a imagem de fundo, é possível utilizar métodos baseados na aparência, pois estes dispensam extração de características ou mesmo pré-processamento dos dados e fazem suas inferências com base na própria imagem (ROTH; WINTER, 2008). Na literatura, métodos baseados na aparência apresentam bons resultados na identificação de padrões em conjuntos de imagens (DELAC; GRGIC; LIATSIS, 2005; KIM; MALLIPEDDI; LEE, 2014). Portanto, identificar o momento onde a cena deixa de representar comportamento normal e passa a conter ações suspeitas, com o uso da abordagem baseada na aparência, pode reduzir o custo computacional da tarefa.

Por outro lado, a seleção de regiões e a marcação de objetos na cena são problemas mais facilmente atacados por métodos baseados em características (KE; SUKTHANKAR; HEBERT, 2007; KE et al., 2013). Eles possibilitam que mais detalhes sejam investigados nos vídeos e podem gerar informações mais completas quanto aos elementos que compõem a cena. Essas informações podem ser utilizadas no processo de reconhecimento do comportamento e podem ajudar a compor as marcações que irão auxiliar o operador humano (SINDHUJA; SRINIVASAGAN; KALAISELVI, 2014). No entanto, métodos baseados em características podem apresentar maior custo computacional em relação aos métodos baseados na aparência.

Este trabalho propõe o uso das duas abordagens em uma arquitetura modular para analisar vídeos de multidão em sistemas de vigilância. Na literatura, combinar métodos baseados em características com métodos baseados na aparência para identificar mudança de comportamento em vídeos não é uma abordagem amplamente investigada, principalmente no que diz respeito ao uso dessas ferramentas em um arquitetura modular que possibilite desativar uma das duas abordagens, no intuito de reduzir o custo do sistema.

Ao utilizar as duas abordagens de forma adequada, o sistema inteligente utilizaria todos os seus módulos somente nos momentos onde um comportamento suspeito está presente na cena. Assim, ao combinar abordagens distintas é possível reduzir o custo da análise do comportamento nos vídeos do sistema de vigilância.

1.3 Objetivos

Esta seção mostra o objetivo geral e os objetivos específicos deste trabalho.

1.3.1 Objetivo Geral

Desenvolver uma arquitetura que combine técnicas baseadas na aparência com técnicas baseadas em características para detectar, em tempo real, comportamento anômalo em cenas de multidão de sistemas de segurança baseados em vídeo.

1.3.2 Objetivos Específicos

- Adaptar a análise do vídeo ao contexto das ações exibidas na cena por meio de uma arquitetura modular onde a avaliação do vídeo seja sensível ao contexto das ações presentes nas cenas e a utilização dos métodos seja feita de forma seletiva, conforme a necessidade de seu uso.

- Validar o uso de métodos baseados na aparência para que estes operem em vídeos de segurança e possam identificar possíveis mudanças de comportamento nas cenas.

- Elaborar uma heurística de limiar adaptativo que atue em conjunto com os métodos baseados na aparência para identificar mudanças de comportamento sem uma etapa de aprendizado.
- Propor o uso do operador de divergência para classificar o comportamento conforme as características intrínsecas do vídeo e gerar marcações que possibilitem aos operadores humanos tomar ciência de provável atividade de risco na cena avaliada.
- Comparar o método proposto com métodos existentes na literatura e avaliar os resultados obtidos com experimentos em vídeos que apresentem mudança de comportamento das pessoas.

1.4 Contribuições

A seguir, são listados os pontos tratados nesta pesquisa que apresentam contribuições na área de Visão Computacional, mais precisamente para o atual estado-da-arte do problema de detectar e reconhecer comportamento anormal em vídeos de multidão.

- **Uma arquitetura modular** com fases independentes que possibilita particionar o problema e reduzir o custo de computação durante a análise dos vídeos.
- **Um método** baseado na arquitetura proposta para detectar e reconhecer comportamento anormal em vídeos de multidão. O método, composto por módulos independentes entre si, mantém apenas parte de seus elementos ativos durante sua execução.
- **Adaptar o uso de métodos baseados na aparência** para operar em vídeos e gerar sinais que possibilitam detectar comportamento anormal em cenas de multidão.
- **Uma heurística de limiar adaptativo** utilizada em conjunto com métodos baseados na aparência para detectar anomalia em vídeos.
- **O uso de um operador de divergência** para avaliar descritores de fluxo óptico e gerar níveis que indicam comportamento anormal de dispersão e de aglomeração em cenas de multidão.

1.5 Estrutura da Dissertação

Este trabalho está organizado da seguinte forma: O Capítulo 2 apresenta alguns fundamentos teóricos para o entendimento do trabalho e o Capítulo 3 descreve os trabalhos relacionados com esta pesquisa. O Capítulo 4 mostra o método proposto e descreve o funcionamento de cada um dos seus módulos. O Capítulo 5 discute como o método foi avaliado e exibe os resultados obtidos nos experimentos realizados. Por fim, as conclusões obtidas e as propostas de trabalhos futuros desta pesquisa são apresentados no Capitulo 6.

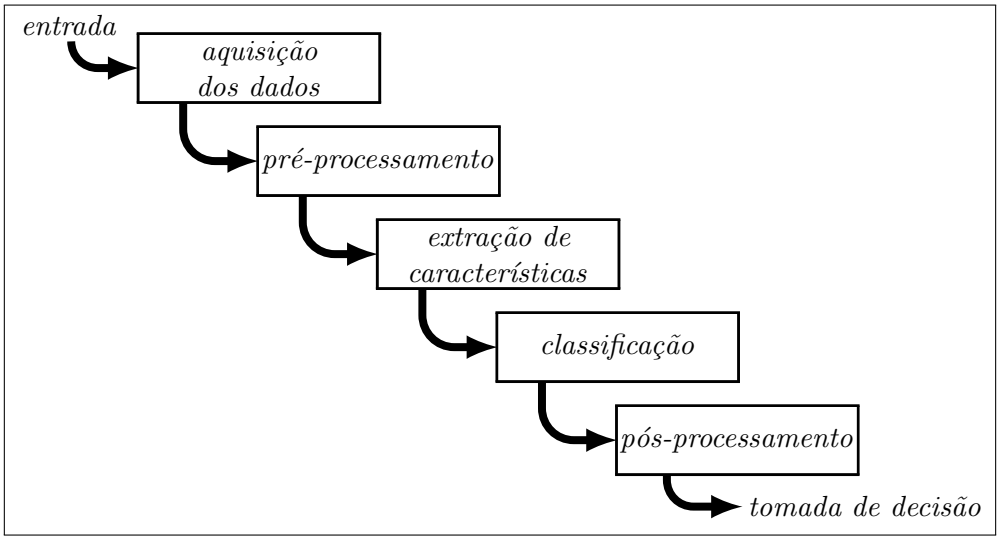
2 Fundamentação Teórica

Este capítulo apresenta alguns conceitos necessários para o entendimento de problemas que envolvem reconhecimento de comportamento em vídeo e mostra como esta tarefa pode ser executada em SVBVs, dado que esse tipo de sistema é o foco deste trabalho. A Seção 2.1 apresenta os fundamentos relacionados aos problemas genéricos de reconhecimento de padrões e descreve as etapas envolvidas no processo. Em seguida, o procedimento geral adotado no reconhecimento de comportamento em vídeos e algumas das técnicas que compõem o estado-da-arte desse tipo de atividade são descritos na Seção 2.2.

2.1 Reconhecimento de Padrões

Muitas das atividade executadas por pessoas no dia-a-dia são atividades complexas, que normalmente envolvem Reconhecimento de Padrões (do inglês, *Pattern Recognition* – PR). Segundo Duda, Hart e Stork (2001), PR busca identificar as categorias de um conjunto de padrões para fazer uso desta informação como base para uma tomada de decisão no contexto de um problema. Porém, para que PR seja aplicado de forma automática em um sistema, é necessário sua divisão em várias etapas. A Figura 1 apresenta o diagrama com as principais etapas presentes no modelo clássico de um sistema de PR. Modelos mais recentes baseados em *Deep Learning* normalmente não seguem esse padrão (PATEL; NGUYEN; BARANIUK, 2015)

Figura 1 – Diagrama de um sistema de reconhecimento de padrões dividido em etapas.



Fonte: Adaptado de Duda, Hart e Stork (2001).

As etapas apresentadas no diagrama podem variar de acordo com o problema abordado, o tipo de dados recebidos na entrada, a tomada de decisão esperada na outra

extremidade ou conforme a natureza do problema e de sua solução, que podem dispensar uma das etapas listadas ou necessitar incluir algum procedimento que não foi descrito. A seguir, são descritas as etapas apresentadas no diagrama da Figura 1.

- *Aquisição dos Dados:* Os dados recebidos como entrada pelo sistema podem ser: coordenadas, características de objetos (formato, tamanho, cor, etc.) ou mesmo um conjunto de imagens. O tipo de dado recebido pelo sistema pode variar de acordo com os sensores utilizados, como um arquivo de texto ou planilha lidos de um disco, uma câmera de vídeo ou sensores que enviam sinais, dentre outros.

- *Pré-processamento:* Esta etapa é responsável pela remoção de ruídos e pela modificação, se necessária, do formato dos dados, para a etapa de extração de características. No caso de um conjunto de imagens, pode ser necessária a aplicação de métodos de segmentação da imagem.

- *Extração de Características:* Nesta etapa é feita a seleção dos dados de interesse presentes no conjunto retornado pelo pré-processamento. Em seguida, os dados extraídos são repassados para o classificador. A extração de características pode ser vista como um filtro que faz a seleção apenas de propriedades relevantes para o classificador. Além disso, selecionar características ruins ou dados ruidosos pode piorar a acurácia do classificador.

- *Classificação:* Etapa onde os objetos são rotulados à uma categoria com base no conjunto de propriedades selecionadas pela etapa de extração de características. A complexidade do processo de rotulação varia conforme o acerto na escolha dos valores analisados na etapa de extração de características. Os níveis de ruído que o conjunto apresenta também influenciam nos resultados gerados na classificação. A escolha do método de classificação deve levar em consideração o modelo utilizado para descrever cada categoria e a variabilidade das suas propriedades em relação às propriedades das outras classes.

- *Pós-processamento:* Ao obter a saída do classificador, a etapa de pós-processamento avalia os resultados e decide por atuar de forma adequada. Caso a acurácia obtida seja baixa, pode ser feito um retorno aos estados anteriores no intuito de modificar o comportamento do sistema como tentativa de aprimorar a acurácia obtida pelo classificador.

Entre as etapas descritas acima, a etapa de extração e/ou seleção das características é uma das mais importantes, pois irá refletir diretamente nos resultados produzidos na etapa de classificação (RASHEED; KHAN; KHALID, 2014). Um sistema para reconhecer o comportamento de pessoas em vídeos de segurança poderá apresentar um conjunto

de etapas semelhante ao apresentado na Figura 1. As etapas que compõem o sistema desenvolvido neste trabalho são apresentadas no Capítulo 4. A próxima seção discute os tipos de métodos encontrados na literatura para reconhecimento de comportamento em vídeos e apresenta conceitos essenciais para o entendimento do contexto deste trabalho.

2.2 Reconhecimento de Comportamento em Vídeos

A utilização de SVBVs em locais como empresas, residências e áreas públicas gera uma grande quantidade de conteúdo em vídeo com ações que podem ser analisadas a fim de identificar comportamentos suspeitos como brigas e assaltos, ou incidentes que ofereçam risco de vida às pessoas, como explosões ou incêndios. Essa análise é normalmente realizada por operadores humanos. Contudo, geralmente o volume de conteúdo produzido é maior do que a quantidade de material que os operadores conseguem analisar de forma adequada, o que torna impraticável que apenas operadores humanos analisem o conteúdo nos vídeos (COHEN et al., 2008; ZIN et al., 2014; HUEBER et al., 2014).

Assim, a utilização de ferramentas capazes de auxiliar operadores humanos na tarefa de avaliar o conteúdo dos vídeos possibilita que uma quantidade maior de dados seja avaliada no intuito de identificar comportamentos suspeitos ou cenários de risco (COHEN et al., 2008; SAINI et al., 2012; POPOOLA; WANG, 2012; BERTINI; BIMBO; SEIDENARI, 2012; ZIN et al., 2014; HUEBER et al., 2014).

Essas ferramentas utilizam técnicas capazes de operar em situações como: variação dos níveis de iluminação, balanço da câmera, mudança do plano de fundo, alto grau de ruído e baixa qualidade do vídeo. Características presentes no vídeo variam de acordo com a cena e com o SVBV utilizado (BERTINI; BIMBO; SEIDENARI, 2012; JODOIN; SALIGRAMA; KONRAD, 2012). Analisar vídeos em busca de comportamento é um processo que possui custo computacional elevado e requer que ações como detecção, reconhecimento e rastreamento de pessoas sejam aplicadas durante a execução do método de reconhecimento de comportamento (SAINI et al., 2012; BOUZEGZA; ELARBI-BOUDIHIR, 2013; SAINI; ATREY; SADDIK, 2014).

Os métodos encontrados na literatura podem ser divididos em dois grupos de acordo com as técnicas utilizadas: (1) Métodos Baseados em Características, que dependem da extração de modelos utilizados para descrever os objetos e suas categorias (DUDA; HART; STORK, 2001; BOUZEGZA; ELARBI-BOUDIHIR, 2013), e (2) Métodos Baseados na Aparência, que são capazes de produzir inferências sem extrair modelos do vídeo e são baseados em métodos estatísticos que utilizam apenas o conteúdo da própria imagem (JIANG et al., 2013; COURTY et al., 2014; ZIN et al., 2014).

A forma como os métodos avaliam as imagens para derivar seus modelos pode variar conforme a natureza do problema ou de acordo com o tipo de método utilizado e podem

ser de escopo global ou local. Quando toda a imagem é utilizada para gerar um modelo matemático, este modelo é dito global. Caso contrário, os modelos são chamados de locais e pode haver mais de um por imagem. Métodos baseados em características são comumente utilizados em aplicações do tipo local, por exemplo, ao marcar uma área de interesse na imagem ou rastrear um objeto na cena. Os métodos baseados na aparência, no entanto, são normalmente aplicados em escopo global, pois avaliam toda a imagem.

A Seção 2.2.1 discute os Métodos Baseados em Características e as principais etapas envolvidas em sua execução. A Seção 2.2.2 apresenta a estrutura geral dos Métodos Baseados na Aparência e descreve o funcionamento de algumas das técnicas capazes de reconhecer comportamentos em vídeos com base na aparência.

2.2.1 Métodos Baseados em Características

Uma das abordagens mais comuns adotada na área de Visão Computacional é a utilização de modelos matemáticos obtidos a partir de imagens ou vídeos. Esses modelos são usados como conjunto principal de dados para realizar tarefas como rastreamento, segmentação e identificação de objetos na cena (BOUZEGZA; ELARBI-BOUDIHIR, 2013). Os dados extraídos podem ser linhas, bordas, conjunto de coordenadas de pontos específicos ou nível de intensidade de iluminação, entre outros, e são tratados como características da imagem, ou do vídeo, capazes de descrever detalhes da cena (CONTE et al., 2010; BOUZEGZA; ELARBI-BOUDIHIR, 2013; LI et al., 2014).

Assim como os sistemas de reconhecimento de padrões, métodos para reconhecer comportamento em vídeos podem ser divididos em etapas. Algumas das principais etapas executados por esses métodos são: reconhecimento do plano de fundo, reconhecimento de objetos e rastreamento em vídeos. Essas etapas são descritas a seguir.

- *Reconhecimento do Plano de Fundo:* Esta é normalmente uma das primeiras tarefas executadas no decorrer do processo de extração das características, pois possibilita que sejam identificados os objetos de interesse na cena (HSU et al., 2013). O método *Background Subtraction* é uma abordagem para reconhecimento do plano de fundo que consiste em definir um modelo para representar a imagem de fundo com base em trechos do vídeo ou em um único quadro inicial. Neste último caso, o modelo pode ser atualizado à medida que os elementos do vídeo são identificados (JODOIN; SALIGRAMA; KONRAD, 2012).

  Outra estratégia utilizada para modelar a imagem de fundo é o Fluxo Óptico (FO), que avalia vizinhanças de quadros para localizar pixels correspondentes e cria um mapa de vetores que descrevem a direção e a velocidade com que cada região do vídeo se deslocou no espaço da imagem no decorrer do tempo. O deslocamento uniforme de várias áreas pode indicar trepidação ou movimento de câmera e permite identificar

o plano de fundo da cena mesmo em vídeos onde a câmera se move (CONG; YUAN; LIU, 2011; KRAUSZ; BAUCKHAGE, 2012).

Em abordagens como Walha, Wali e Alimi (2013) e Zin et al. (2014), a modelagem da imagem de fundo interfere diretamente no reconhecimento do comportamento, pois a segmentação e o rastreamento dos objetos na cena dependem da qualidade dos resultados e da robustez do método de reconhecimento do plano de fundo (POPOOLA; WANG, 2012).

- *Detecção de Objetos:* Vários problemas na área de visão computacional estão relacionados com a detecção de objetos em vídeos. Neste caso, um objeto pode ser: uma pessoa, um grupo de pessoas, veículos e malas que são abandonadas. Esta tarefa se baseia em modelos que descrevem as características dos objetos como: velocidade, localização, volume ou trajetória pela linha de tempo do vídeo (BOUZEGZA; ELARBI-BOUDIHIR, 2013). Como o foco deste trabalho envolve cenários que possam conter multidão, isto é, duas ou mais pessoas presentes ao mesmo tempo na cena, não é viável que o método seja baseado na contagem de pessoas na imagem, o que pode acarretar em um número máximo de pessoas suportado pelo método, dado o custo computacional apresentado (LI et al., 2014). Assim, este trabalho irá tratar apenas modelos que avaliam o conjunto e não os indivíduos que compõem a multidão.

- *Rastreamento em Vídeos:* Uma tarefa comum ao reconhecer comportamento em vídeos é rastrear os objetos detectados na cena. O rastreamento normalmente é realizado com foco específico no objeto, mas é possível que este seja aplicado a uma região, como por exemplo, um grupo de pessoas que se desloca na cena (ZHAO; LI, 2014). Esse tipo de abordagem normalmente é utilizado em métodos de reconhecimento de comportamento que avaliam o fluxo, como o FO. Uma das vantagens deste tipo de abordagem é que a quantidade de pessoas na cena não representa um problema para o algoritmo, que pode ter apenas uma pessoa ou uma multidão (LI et al., 2014).

  Em abordagens que avaliam a trajetória, é comum a seleção dos percursos conforme sua semelhança em relação aos caminhos pré-rotulados, normalmente utilizada quando se tem um conjunto de modelos pré-definidos para a cena (CHONG et al., 2014). É possível que a identificação seja feita por características que descrevem o contexto e que descrevem a aparência.

A extração de características baseada no plano de fundo é uma atividade que possui elevado custo computacional, pois avalia a imagem em nível de pixel e compara cada ponto com seus vizinhos (JODOIN; SALIGRAMA; KONRAD, 2012). Quando os métodos são aplicados em vídeos, são feitas avaliações entre pixels de quadros vizinhos ou entre conjuntos de quadros (ZIN et al., 2014; WALHA; WALI; ALIMI, 2013; BOUZEGZA; ELARBI-BOUDIHIR, 2013).

Existem métodos que não utilizam informações do plano de fundo, como Guodong (2011), Bertini, Bimbo e Seidenari (2012) e Walha, Wali e Alimi (2013), eles utilizam algoritmos para extração de características independentes quanto à escala, rotação e translação da imagem ou buscam por padrões periódicos identificados quando o vídeo é analisado no âmbito temporal e descrevem o comportamento como uma linguagem natural.

Uma abordagem baseada em FO pode dispensar etapas como pré-processamento ou reconhecimento do plano de fundo, pois é possível criar um modelo do comportamento da multidão que, como pode ser visto em Li et al. (2014), é baseado em continuidade. O FO permite tratar a multidão como um conjunto de partículas que flui pela cena, que pode se contrair, expandir ou dissipar. Esse método é descrito em detalhes na próxima seção.

2.2.1.1 Fluxo Óptico

O FO é um método que estima o fluxo do deslocamento de objetos nas imagens e constrói um mapa de vetores que representa velocidade e sentido do fluxo dos objetos na cena. Este movimento é referente ao deslocamento no plano bidirecional da imagem e é uma projeção do movimento dos objetos no plano tridimensional da cena (BEAUCHEMIN; BARRON, 1995; YILMAZ; JAVED; SHAH, 2006). O método, na sua forma mais simples, funciona da seguinte forma: dado um quadro Q de um vídeo no tempo t e um mapa de fluxo $\mathbf{x}$ – com $\mathbf{x}$ para representar conjunto de vetores (u, v) no plano da imagem – para Q, o cálculo do fluxo ente $Q^{\,t}_{\mathbf{x}}$ e o próximo quadro $Q'^{(t+1)}_{\mathbf{x}'}$ é dado por:

$$Q'^{(t+1)}_{\mathbf{x}'} \approx T(Q', Q^t_{\mathbf{x}}) \tag{2.1}$$

com o fluxo para $\mathbf{x}'$ obtido a partir da transformação T onde:

$$\mathbf{x}' = \Gamma_{\mathbf{x}} + \Lambda_{\mathbf{x}}\mathbf{x}\varepsilon. \tag{2.2}$$

Na Equação (2.2), $\Gamma_{\mathbf{x}}$ é a transformação de translação aplicada a $\mathbf{x}$, $\Lambda_{\mathbf{x}}$ é a rotação e ε, um escalar. Os valores de $\mathbf{x}$ referentes ao fluxo de Q no tempo t representam a velocidade de um ponto $\mathbf{v}$ da imagem Q dada por:

$$\mathbf{v} = \frac{\delta \mathbf{x}}{\delta t} \tag{2.3}$$

com distância δ entre cada um dos valores de $\mathbf{x}$ e de t dos quadros (BERNARD, 1999).

Este modelo apresentado é a abordagem básica do FO e considera que a iluminação da cena é constante. O modelo de superfícies Lambertianas[1] é adotado por tornar possível a aplicação do FO em vídeos com variação de iluminação, cenário predominante em vídeos

[1] Uma superfície Lambertiana reflete toda a luz incidente sobre ela e apresenta brilho uniforme para todas as direções que a superfície é visualizada.

reais, conforme a equação:

$$Q'^{(t+1)}_{\mathbf{x}'} \approx T(Q', Q^t_{\mathbf{x}}) P(Q', Q^t_{\mathbf{x}}) \tag{2.4}$$

em que a transformação P é referente à iluminação da cena que, em um modelo simples, pode ser representado por:

$$P(Q', Q^t_{\mathbf{x}}) = \frac{Q}{d^2} \cos(\theta) \tag{2.5}$$

onde θ é o ângulo entre a superfície do objeto e a fonte de luz da cena e d é a distância entre eles.

Na literatura são encontrados vários métodos baseados no FO que apresentam formas diferentes de obter dados do fluxo ou que direcionam o método para solução de outros problemas como identificar e rastrear objetos na cena. Movimentos de câmera, ruídos, mudanças de iluminação na cena e elementos que fazem parte do plano de fundo, mas que apresentam algum movimento, são tratados com métricas que definem um limiar para o tamanho dos vetores de descrição do fluxo ou sentido do deslocamento. Essas métricas possibilitam tratar o ruído tanto no quesito temporal quanto no aspecto espacial.

2.2.2 Métodos Baseados na Aparência

Uma abordagem muito utilizada para reconhecer padrões em conjuntos de imagens são os métodos baseados em modelos matemáticos gerados a partir dos níveis de intensidade para toda a imagem ou de uma determinada região da cena. Estes modelos, baseados na aparência, podem ser conjuntos estatísticos simples, como a média dos valores de intensidade dos pixels, níveis de histogramas ou valores obtidos ao aplicar técnicas de redução de dimensionalidade, como Análise dos Componentes Principais ou Análise dos Componentes Independentes (ROTH; WINTER, 2008). Normalmente, métodos baseados na aparência utilizam modelos derivados para toda a imagem, onde a abordagem é dita global.

Para a criação dos modelos matemáticos, existem estratégias que apresentam menor custo computacional ao fazer a redução de dimensionalidade dos dados, tendo em vista que, para uma abordagem global, vídeos com dimensões maiores irão acarretar em alto tempo de resposta do método. Nas próximas seções são apresentados os métodos Análise dos Componentes Principais (do inglês, *Principal Component Analysis* – PCA), PCA Incremental (do inglês, *Incremental PCA* – IPCA), PCA bidirecional (do inglês, *Bi-directional PCA* – 2D-PCA) e o Método do Subespaço Mútuo (do inglês, *Mutual Subspace Method* – MSM), chamados de métodos do subespaço pois projetam a imagem em um sistema de coordenadas com dimensionalidade reduzida e mantêm os dados que apresentam maior relevância para a atividade realizada (ROTH; WINTER, 2008). Esse conjunto de dados com menor dimensionalidade é a projeção dos dados no subespaço. Assim, com a projeção da imagem original em um subespaço, é gerada uma nova representação para os dados.

São apresentados dois outros métodos baseados na aparência, mas que não são métodos do subespaço: o Índice de Similaridade Estrutural (do inglês, *Structural Similarity Index* – SSIM), métrica criada para medir a qualidade de uma imagem e o método da Distância Euclidiana entre Imagens (do inglês, *Image Euclidean Distance* – IMED), que mede a semelhança entre duas imagens ao avaliar a diferença entre cada pixel da imagem.

Ao tratar problemas que envolvem casamento de padrões, abordagens baseadas nos métodos do subespaço avaliam seus resultados ao medir a distância ou, no caso do MSM, o ângulo canônico entre as representações construídas por ele. Projeções próximas umas das outras indicam maior semelhança entre as imagens originais, pois os subespaços gerados são uma maneira de representar a estrutura das imagens de forma otimizada (GATTO; HINO; FUKUI, 2014). O método IMED mede a distância entre imagens sem gerar uma representação, enquanto que o SSIM deriva um grau de relação entre duas imagens com base apenas na similaridade de suas estruturas. Os métodos são descritos com mais detalhes nas próximas seções.

2.2.2.1 Análise dos Componentes Principais

A Análise dos Componentes Principais é uma técnica amplamente utilizada em áreas como reconhecimento de padrões e visão computacional para redução da dimensionalidade dos dados (CHOI et al., 2011; KIM et al., 2013). O método PCA projeta os dados originais, conforme sua variância, em um subespaço que contém um conjunto menor de elementos para avaliação. O PCA é uma projeção ortogonal que utiliza os auto valores da matriz de covariância para determinar os componentes principais. Existem várias formas de se aplicar o PCA, no entanto, para as áreas de visão computacional e classificação de imagens foram feitas modificações no método original para que sejam considerados a estrutura da imagem e o tempo gasto pelo método, dadas as dimensões das imagens analisadas.

Na abordagem clássica do método PCA, a imagem I como dado de entrada com H pixels de largura e V pixels de altura é transformada em um vetor com dimensão $H \times V$. Ao aplicar o PCA diretamente como um classificador baseado em imagens, com um conjunto S_n de n imagens vetorizadas, uma nova imagem $\mathbf{O}$ pode ser classificada ao avaliar sua correspondência com os elementos rotulados disponíveis na base. Com S e $\mathbf{O}$ normalizados, $||S_n|| = ||\mathbf{O}|| = 1$, portanto:

$$||S_n - \mathbf{O}||^2 = 2 - 2\ S_n^{\mathsf{T}}\ \mathbf{O} \tag{2.6}$$

e podemos nos basear na soma do quadrado das diferenças como critério simplificado de classificação e utilizar um limiar ψ para determinar a classe do novo objeto, com $||S_n - \mathbf{O}||^2 > \psi$.

Aplicar o PCA diretamente aos valores de intensidade dos pixels apresenta desvantagens quanto ao grande número de pixels nas imagens que acarreta no aumento do custo computacional e quanto às variações nos níveis de iluminação na cena que não sejam referentes ás mudanças na estrutura da imagem, que podem degradar o modelo gerado. O método PCA é normalmente utilizado em grandes conjuntos de dados para gerar um subespaço com dimensão reduzida e a classificação é feita com base nas projeções geradas.

2.2.2.2 Análise dos Componentes Principais Incremental

Ao trabalhar com conjunto de imagens, o PCA pode não ser capaz de operar com toda a informação recebida como entrada, devido sua abordagem ser acumulativa. O conteúdo recebido é utilizado na construção da nova representação, mas uma quantidade muito grande de informação torna inviável processar os dados devido ao tempo necessário e a limites de memória. Para solucionar esse problema, pode ser aplicada uma abordagem incremental do método PCA, como a proposta por (CHANDRASEKARAN et al., 1997).

Nessa abordagem, o subespaço é modificado sem recalcular toda a matriz de covariância, o que reduz a quantidade de memória necessária para manter o método em constante atualização, uma vez que os dados obtidos anteriormente não são necessários para a atualização do modelo. O tempo da atualização reduz, pois a decomposição em valores singulares (do inglês, *Singular Value Decomposition* – SVD) é aplicada a um conjunto menor de dados (ZHAO; YUEN; KWOK, 2006).

O comportamento do IPCA é semelhante ao PCA. Ele difere na forma como o novo elemento é inserido no subespaço. A atualização do modelo atual Y_i gerado para i imagens é dada pelo cálculo aproximado da SVD apenas para a nova imagem Y_{i+1}. Sejam U_i e V_i matrizes do modelo atual que contêm, respectivamente, os k vetores singulares mais à esquerda e mais à direita, onde k é o tamanho do subconjunto $t - rank$ da nova imagem Y_{i+1} para um limiar de tolerância t. O valor de t pode ser definido com base nos menores valores singulares o_1 e o_2 de Y_{i+1}, onde $o_1 > t >= o_2$ e t é sempre muito menor que i. Seja S_i uma matriz diagonal quadrada de dimensão k, com os valores singulares, a atualização pode ser feita com:

$$[\ U_i \ \ S_i \ \ V_i^{\mathsf{T}} \ \ Y_{i+1}\ ] \ = \ U' \ S' \ V'^{\mathsf{T}} \tag{2.7}$$

onde U e V terão atualizações somente para as k primeiras colunas de U' e V', respectivamente. Para S, a atualização irá ocorrer somente para a submatriz com as primeiras k linhas e k colunas de S'. Assim, os novos valores que irão compor o modelo são uma estimativa da SVD para a imagem Y_{i+1}.

2.2.2.3 Análise dos Componentes Principais Bi-direcional

Existem abordagens de PCA que não convertem os dados para vetores, e os mantém na forma de matriz. Esses métodos são chamados de PCA bidimensional (2D–PCA) e possibilitam preservar a relação espacial do conteúdo presente na imagem (SANGUANSAT, 2010). Ao manter as imagens como matrizes, é possível construir uma matriz de covariância para as imagens muito menor que a matriz de covariância obtida no PCA convencional (YANG et al., 2004). Como resultado, o 2D–PCA tem maior velocidade de execução sem prejudicar a acurácia do método. Além disso, no método PCA convencional cada componente principal obtido é um escalar, enquanto que no 2D–PCA o componente principal é um vetor, que compõe a matriz de características da imagem, chamada de imagem de características.

Dada a projeção P de uma imagem Y de tamanho $m \times n$ em X, por:

$$P = Y\ X' \tag{2.8}$$

onde X é um vetor de dimensão n, para que o cálculo da matriz de covariância no 2D–PCA resulte em uma boa projeção da imagem no espaço, é adotado o seguinte critério:

$$C(X) = traço(X^{\mathsf{T}}\ G\ X). \tag{2.9}$$

A função $traço(\cdot)$ retorna a soma dos elementos da diagonal principal de uma matriz. O elemento G na Equação (2.9) é a matriz de covariância dispersa dada por:

$$G = \frac{1}{N}\sum_{i}^{N}(Y_i - \bar{Y})^{\mathsf{T}}\ (Y_i - \bar{Y}) \tag{2.10}$$

para um conjunto Y_i com N imagens, onde $i = \{1,\ 2,\ 3,\ ...\ ,\ N\}$ e $\bar{Y}$ é a imagem média de Y_i.

Uma boa projeção dada por X é o vetor que maximiza o critério estabelecido na equação (2.9). São construídas N projeções, as quais formam o conjunto X_i. Estas projeções são utilizadas no processo de extração das características que formam a imagem de característica. Para as N imagens, teremos então um conjunto Pi com N projeções para uma única imagem Y de Y_i, dado por:

$$Pi = Y\ Xi \tag{2.11}$$

onde $i = \{1,\ 2,\ 3,\ ...\ ,\ N\}$ e P_i é a imagem de características para Y.

O uso do método 2D–PCA para reconhecimentos de padrões se dá pela avaliação da distância entre as representações geradas. A medida mais comum é a distância euclidiana, e seu uso apresenta bons resultados diante da maioria dos métodos baseados no PCA e de outros métodos do subespaço (FAWZY; ABDELWAHAB; MIKHAEL, 2013; KIM; MALLIPEDDI; LEE, 2014).

2.2.2.4 Método do Subespaço Mútuo

O Método do Subespaço Mútuo, uma extensão do Método do Subespaço, assim como os métodos apresentados anteriormente, aplica uma redução da dimensão dos dados de entrada. Segundo Gatto, Hino e Fukui (2014), o MSM apresenta bons resultados no processo de reconhecimento de padrões em conjuntos de imagens, pois consegue armazenar, de forma implícita, a estrutura dos objetos recebidos para o aprendizado. O método do subespaço mútuo consiste dos mesmos passos do método do subespaço, exceto pela forma como a entrada é construída. No MSM, a entrada é um subespaço e não mais um único vetor. Essa peculiaridade permite analisar a similaridade entre conjuntos de padrões, os quais são espaços com dimensão reduzida.

O processo de redução da dimensão consiste de uma transformação que, aplicada à uma imagem de entrada I_d vetorizada com dimensão d, gera um vetor de características x_k com dimensão k, onde $k \ll d$. No método MSM, é construído um subespaço para cada novo conjunto de imagens. Os subespaços são gerados pela extração de k auto vetores da matriz de correlação M, com:

$$M = \frac{1}{n} \sum_{i=1}^{k} (x_i)(x_i)^{\mathsf{T}} \tag{2.12}$$

que será um conjunto vetorizado de imagens. No MSM, para obter a matriz M não é realizado o cálculo da média dos valores, como é feito no método PCA.

O método MSM é baseado em cálculo de ângulos canônicos e aplica em um de seus passos o PCA. O método explora a possibilidade de condensar um conjunto de imagens em uma representação eficiente na forma de grupos de vetores ortonormais, e forma assim um subespaço com dimensão reduzida. Ao selecionar os valores singulares de M, é possível extrair os ângulos canônicos entre os subespaços. Este ângulo representa a similaridade estrutural entre as imagens que originaram o subespaço.

2.2.2.5 Índice de Similaridade Estrutural

Ao levar em consideração a importância da similaridade estrutural das representações geradas para os métodos do subespaço, foi incluído neste trabalho o método Índice de Similaridade Estrutural (SSIM). O SSIM é um método criado para gerar uma medida de similaridade baseada na sensibilidade da mudança da estrutura da imagem como um todo. Esta característica está presente no sistema visual humano (DHALL; ASTHANA; GOECKE, 2010).

A análise de similaridade estrutural entre imagens possibilita avaliar a relação entre a vizinhança de um dado trecho da imagem diante de um elemento em foco, como uma linha, um contorno ou mesmo um objeto presente na imagem (WANG et al., 2004). O método SSIM calcula a similaridade entre duas imagens x e y em escala de intensidade por meio

da seguinte fórmula:

$$ISE(x,y) = \frac{(2\mu_x\mu_y + c_1) \times (2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1) \times (\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2.13}$$

onde σ_{xy} é a covariância entre as imagens, c_1 e c_2 são variáveis para normalizar o cálculo da similaridade com base no raio de influência dos pixels, que varia de 0 a 255, e μ_i e σ_i são, respectivamente, a média e a variância da imagem i. Os valores de similaridade obtidos através da Equação (2.13) são normalizados, onde 1 é a similaridade máxima entre as imagens e zero é a mínima. Assim, aplicar o SSIM para a mesma imagem em x e y retorna o valor 1.

2.2.2.6 Distância Euclidiana entre Imagens

A distância euclidiana é a medida de distância utilizada pela maioria dos métodos do subespaço. As projeções no subespaço são classificadas conforme a distância entre cada uma de suas instâncias. Como forma de avaliar a viabilidade da criação de uma representação para as imagens, o método Distância Euclidiana entre Imagens (IMED) foi investigado. A abordagem clássica do IMED compara os pixels da imagem onde, para duas imagens x e y com dimensões $M \times N$, a distância é calculada por:

$$d(x,y) = \sqrt{\sum_{i}^{M}\sum_{l}^{N}[(x_i - y_i)^2 + (x_l - y_l)^2]} \tag{2.14}$$

que percorre toda a matriz de pixels do par de imagens. Normalmente o IMED é aplicado devido sua simplicidade, pois não gera uma nova representação para as imagens (WANG; ZHANG; FENG, 2005). A distância euclidiana entre imagens leva em consideração a relação entre os valores dos pixels da imagem, mas não se preocupa com a representação dos objetos presentes na figura. Como resultado, o IMED apresenta alta sensibilidade a pequenas deformações da imagem, o que resulta em uma elevada distância mesmo entre imagens visualmente próximas. Isso ocorre por que o método não leva em consideração que as matrizes recebidas como entrada são imagens, e que existe relação entre o pixel e seus vizinhos (WANG; ZHANG; FENG, 2005; CHEN et al., 2006).

Os métodos apresentados neste capítulo são algumas das principais ferramentas utilizadas para reconhecimento de padrões que compõem ou servem de base para os principais métodos estado-da-arte na análise de imagens e vídeos baseada em técnicas de visão computacional. No próximo capítulo, são apresentados alguns dos trabalhos presentes na literatura que se relacionam com esta pesquisa. Suas principais características e desvantagens são discutidas e ao final do capítulo, uma breve comparação é apresentada onde as abordagens são comparadas com o método proposto por esta pesquisa, descrito em maiores detalhes no Capítulo 4.

3 Trabalhos Correlatos

O reconhecimento de comportamento baseado em vídeo é um problema cada vez mais investigado na área de pesquisa de Visão Computacional, motivado pelo grande número de problemas reais que podem ser investigados, geralmente relacionados com segurança, tais como: identificar intrusos ou arruaceiros, reconhecimento de atividades danosas ou suspeitas e comportamento de multidões que possa gerar situações de risco, como pânico generalizado, esmagamento e brigas (KRAUSZ; BAUCKHAGE, 2012; BOUZEGZA; ELARBI-BOUDIHIR, 2013; ZIN et al., 2014; CHONG et al., 2014; HUEBER et al., 2014). A discussão deste trabalho prioriza problemas que envolvem cenas adquiridas por SVBVs.

Este capítulo descreve alguns dos principais trabalhos relacionados com esta pesquisa. A Seção 3.1 trata de alguns dos trabalhos que atacam problemas específicos como reconhecer interação entre pessoas ou identificar objetos abandonados. Na Seção 3.2, são apresentados métodos que visam reconhecer comportamento em vídeos de segurança, mas sem foco em cenas com multidão. Os trabalhos voltados para reconhecer comportamento em vídeos de multidão são discutidos na Seção 3.3. Por fim, a Seção 3.4 discute, de forma breve, os conceitos abordados neste capítulo e os métodos apresentados são comparados.

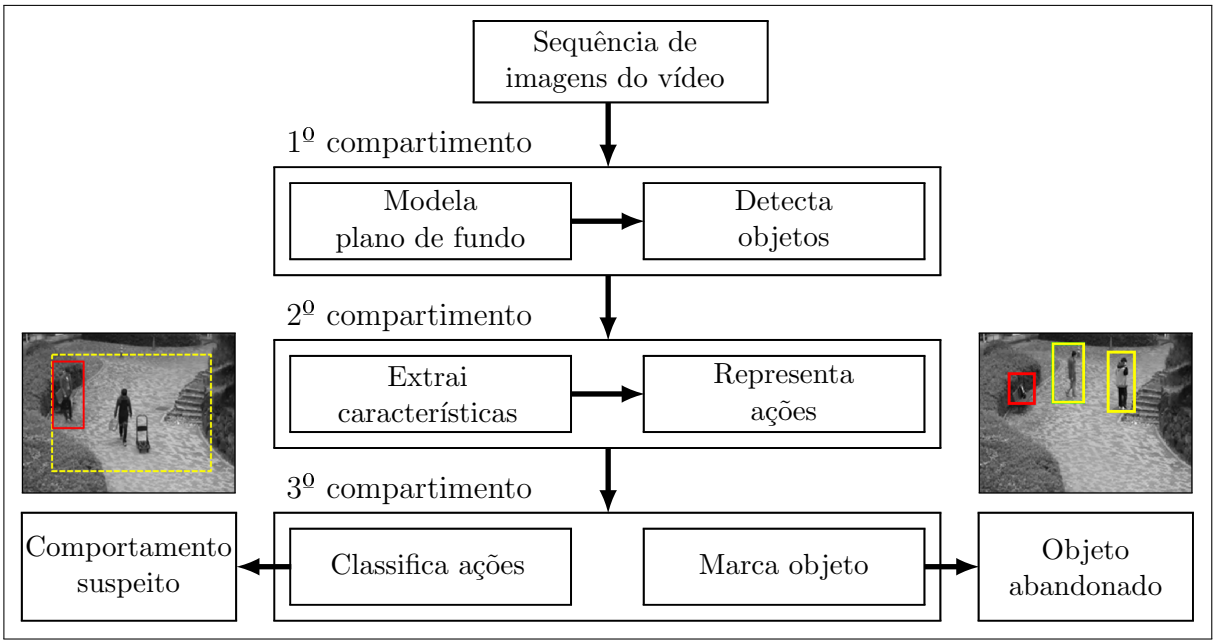
3.1 Aplicações específicas

A seguir, são descritos alguns dos trabalhos relacionados com esta pesquisa que tratam aplicações específicas que envolvem segurança, reconhecimento de comportamento ou extração de características de vídeos.

Conforme mencionado no capítulo anterior, os métodos que trabalham com vídeos podem avaliar a cena no escopo global ou local. Na análise global, a imagem completa é rotulada com o comportamento identificado, o que possibilita marcações na linha de tempo do vídeo. Na análise local, os rótulos são atribuídos à regiões da cena onde o comportamento é identificado. O trabalho apresentado por Zin et al. (2014) propõe uma arquitetura que aplica as abordagens globais e locais em momentos distintos para reconhecer três comportamentos anormais: roubo, brigas e objetos abandonados. O método é dividido em três etapas principais, chamadas compartimentos, como mostra a Figura 2.

O primeiro compartimento, de escopo global, é responsável por modelar o plano de fundo e identificar objetos abandonados. Em seguida, o segundo compartimento, que atua em escopo local, utiliza as características que descrevem aparência, movimento e trajetórias geradas pelo primeiro compartimento para gerar uma Cadeia de Markov Embutida (do inglês, *Embedded Markov Chain* – EMC), modelo probabilístico utilizado para descrever eventos recorrentes que envolvem processos estocásticos. As probabilidades das

Figura 2 – Arquitetura proposta por (ZIN et al., 2014).



Fonte: Adaptado de Zin et al. (2014).

transições da EMC descrevem a correspondência entre as características e os estados representam objetos e pessoas. O terceiro compartimento, responsável por identificar o comportamento na cena, avalia a sequência com que os estados da EMC são alcançados e marca o tempo e a região no espaço da imagem onde o comportamento ocorre, conforme a ação é identificada.

Ao identificar interação entre duas ou mais pessoas, o comportamento pode ser rotulado como uma briga ou roubo. Quando a sequência de ativação da EMC indica interação entre uma pessoa e um objeto, a cena é marcada como abandono de objeto. Uma desvantagem desta proposta é o custo de computação da EMC, que está relacionado com a quantidade de pessoas e objetos na cena. Quanto maior o número de objetos, maior é o custo de computação da EMC. Dessa forma, a abordagem apresentada por Zin et al. (2014) pode não ser viável, do ponto de vista computacional, em cenas que apresentem dezenas de pessoas ou objetos. Além disso, o método não possibilita desativar partes da arquitetura durante sua execução. Assim, todos os seus compartimentos são executados mesmo que haja somente o plano de fundo na sequência de vídeo.

Modelos de escopo local e global podem ser combinados para aprimorar o resultado gerado pelo sistema, como mostra a abordagem desenvolvida por Jiang et al. (2011). A pesquisa é voltada para o problema de reconhecer trajetórias anormais de pessoas e veículos. Ao avaliar a cena no escopo local, o método combina cada trajetória com um modelo global do comportamento da cena. O método cria modelos que descrevem trajetórias criadas por rastreadores de objetos baseados em *Background Subtraction*. O modelo é rotulado e as trajetórias normais na cena são aprendidas por um Modelo Oculto de Markov

(do inglês, *Hidden Markov Model* – HMM), modelo estatístico utilizado para representar processos estocásticos complexos. O HMM representa um conjunto de coocorrências e as trajetórias que não se adequam ao modelo existente são rotuladas como anormais. Um ponto fraco dessa abordagem é que para cada cena deve ser criado um novo modelo de trajetórias.

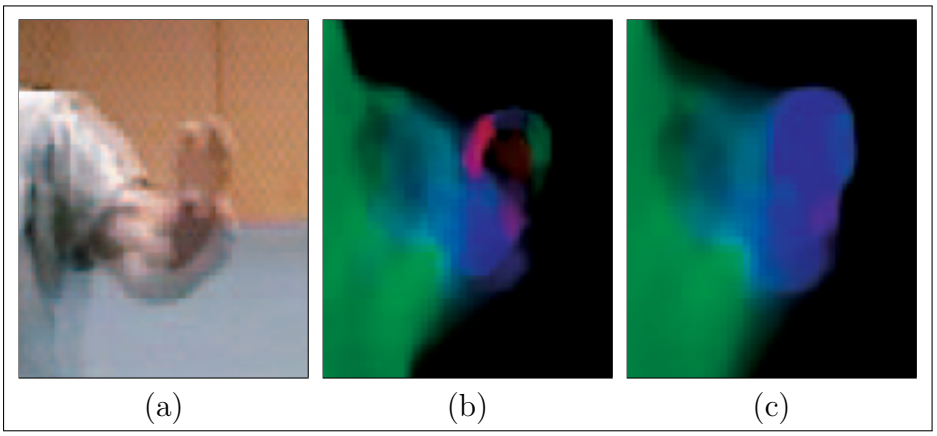
Métodos que não dependam totalmente dos modelos que descrevem o comportamento em uma cena específica são uma vantagem em aplicações reais, pois é normal que mudanças aconteçam em vídeos de sistemas de vigilância. Um exemplo de método capaz de usar um único modelo para várias cenas é a proposta apresentada por Jiang et al. (2013). Essa pesquisa é voltada para reconhecer movimentos de queda em um conjunto de cenas. Vídeos com uma série de simulações do mesmo tipo apresentam movimentos de queda aprendidos por um HMM.

Neste caso, o mesmo modelo pode ser utilizado em vários cenários para identificar queda, com restrições apenas quanto ao ângulo do enquadramento da pessoa na cena. Para rotular as poses, Jiang et al. (2013) utilizam o classificador Máquina de Vetores de Relevância, técnica de aprendizagem de máquina baseada em inferência Bayesiana. Os autores utilizaram esse classificador por ele utilizar menos funções de base que o classificador Máquina de Vetores de Suporte (do inglês, *Support Vector Machine* – SVM). Uma desvantagem do método é sua necessidade de um bom rastreador para a cena, pois o método necessita do modelo local adequado da pessoa para identificar sua pose. Além disso, a pesquisa não trata problemas em cenas com multidão, e sim, cenários onde somente uma pessoa está presente na imagem.

Na literatura, o FO é a abordagem comumente utilizada para avaliar cenas de multidão. Ao utilizar o FO, o número de pessoas no vídeo não representa um problema quanto ao custo de computação, uma vez que seus descritores de movimento não dependem do número de objetos na imagem. Como o FO cria vetores de fluxo com intensidades que refletem a velocidade do movimento na cena, é possível que parâmetros sejam ajustados para que a escala dos vetores se adeque à uma cena específica. Porém, cenas que tendem a variar o ritmo dos movimentos podem não ter um valor ótimo para toda a sequência de movimentos o vídeo. Nesse contexto, Brox e Malik (2011) apresentam uma proposta para calcular o fluxo óptico nesse tipo de vídeo, o *Large Displacement Optical Flow* (LDOF).

O trabalho foca em cenas com oclusão ou movimentos rápidos. Com base em técnicas de suavização de contorno e seleção de pixels, descritores são combinados para que a energia seja captada de forma adaptativa. O modelo que descreve o deslocamento é atualizado para que o fluxo seja condizente com o movimento. A Figura 3 (a) mostra um exemplo de movimento rápido, 3 (b) como um método convencional estima o fluxo e 3 (c) como o método proposto por Brox e Malik (2011) identifica o fluxo na mesma cena. Neste último caso, o movimento da mão não é perdido.

Figura 3 – (a) movimento rápido na cena, (b) fluxo estimado por um método convencional e (c) fluxo obtido pelo método LDOF na mesma cena.



Fonte: Adaptado de Brox e Malik (2011).

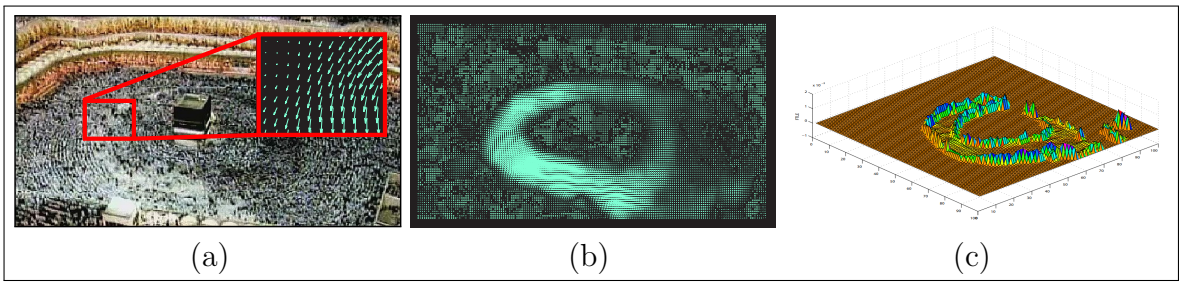
A correspondência de descritores é utilizada como otimização para detectar os deslocamentos na cena e a energia captada é tratada como uma distribuição Laplaciana, onde as variáveis envolvidas são consideradas interindependentes diante da distribuição dos valores observados. Os descritores do FO podem ser utilizados para descrever comportamento no escopo global e local, mas seu alto custo de computação é uma desvantagem, visto que seu uso em vídeos, normalmente, requer que o FO seja derivado para cada par de quadros e em toda a imagem, ainda que grande parte da cena não apresente qualquer movimento.

Uma forma de reduzir o custo de computação em métodos como o descrito por Brox e Malik (2011), é utilizar uma abordagem que não avalie áreas do plano de fundo. No entanto, modelos que descrevem o plano de fundo com baixa qualidade podem propagar erros para as demais etapas do método. Nesse contexto, a pesquisa apresentada por Seo e Kim (2014) propõe um método para modelar o plano de fundo. A proposta apresentou bons resultados ao utilizar uma extensão do PCA, chamada 2D^2-PCA, para criação do modelo que descreve o plano de fundo. Nesse método, um mapa de limiares é definido e cada pixel pode ser rotulado como fundo ou objeto. Como o método é voltado para SVBVs, os modelos que descrevem a imagem de fundo são atualizados no intuito de tratar casos como: variações na iluminação, movimento da vegetação e objetos levados pelo vento, como folhas, papéis e sujeira em geral. O método apresentou bons resultados tanto em cenas externas quanto em ambientes internos, mas sempre com taxas inferiores a 15 quadros por segundo, uma desvantagem do método, principalmente em aplicações reais de vigilância.

Trabalhos como Allain, Courty e Corpetti (2012) e Li et al. (2014) relacionam pesquisas que tratam cenas de multidão como uma entidade contínua. Essa abordagem é adotada no trabalho desenvolvido por Ali e Shah (2007), que apresenta uma proposta para segmentar cenas de multidão. A abordagem assume que o movimento aparente do fluxo óptico derivado tem comportamento semelhante a um conjunto de partículas que

fluem no espaço da imagem. O método aplica um operador de divergência para gerar um modelo que descreve o deslocamento da multidão no tempo. A Figura 4 (c) exibe o resultado obtido por Ali e Shah (2007) ao calcular a Divergência de Lyapunov para o fluxo da Figura 4 (b). A Divergência de Lyapunov mede a taxa de convergência/divergência entre partículas vizinhas e maximiza a região na forma de cumes. Os cumes são tratadas como bordas que separam segmentos de fluxo com dinâmica distinta. A Figura 4 (a) mostra um dos quadros do vídeo e destaca os descritores do fluxo óptico para um trecho da cena.

Figura 4 – (a) Quadro do vídeo onde foi computado o mapa do fluxo, no destaque o fluxo para uma região da cena, (b) mapa completo do fluxo estimado na cena e (c) níveis de divergência/convergência para o fluxo estimado.



(a) (b) (c)

Fonte: Adaptado de Ali e Shah (2007).

Os contornos extraídos do operador de divergência são utilizados para dividir a cena em segmentos. O método combina informações como: sentido do deslocamento, velocidade dos movimentos, tamanho e quantidade de segmentos na cena para marcar pontos de instabilidade nos movimentos da multidão. Uma desvantagem dessa abordagem é sua dependência dos contornos para identificar um região de instabilidade. O método pode não produzir resultados caso a movimentação das pessoas não apresente segmentos por parte do operador de divergência, mesmo que haja mudança dos movimentos na cena.

A Tabela 1 exibe os métodos citados nesta seção e compara algumas de suas características.

Tabela 1 – Métodos descritos que tratam aplicações específicas, comparados quanto ao suporte a cenas de multidão, tipo de abordagem (global/local) e classificador.

Autores	Suporta Multidões	Avaliação Global/Local	Classificador
Ali e Shah (2007)	Sim	Global	–
Brox e Malik (2011)	Sim	Global	–
Jiang et al. (2011)	Sim	Global	MOM e vizinhos mais próximos
Jiang et al. (2013)	Não	Local	HMM e SVM
Zin et al. (2014)	Não	Global e Local	Cadeia de Markov
Seo e Kim (2014)	Não	Global	$2D^2$–PCA

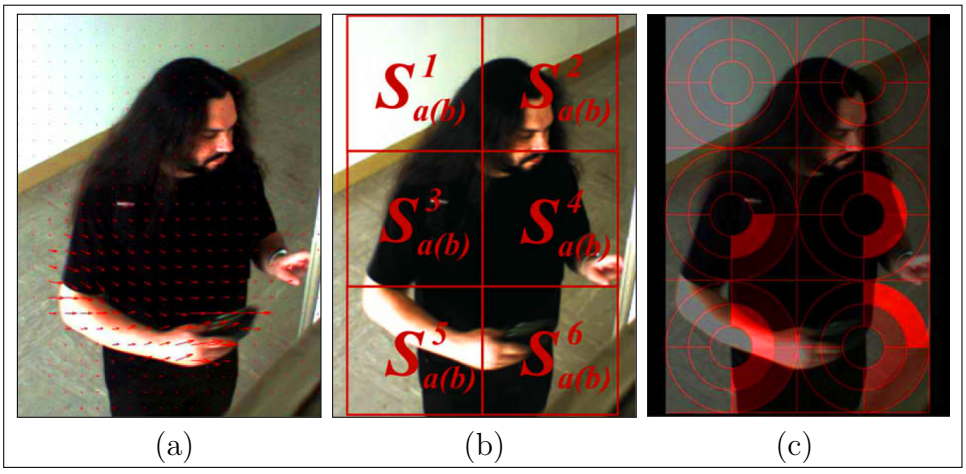
Fonte: O autor(2016).

3.2 Detecção de comportamento anômalo em vídeos sem multidão

Identificar um comportamento em um vídeo muitas vezes requer que informações extraídas da cena sejam combinadas, tais como velocidade dos movimentos, postura das pessoas e o tamanho da área onde a ação ocorre (JODOIN; SALIGRAMA; KONRAD, 2012). Alguns métodos, no entanto, focam em cenários onde somente um pessoa está presente na cena, para identificar comportamentos como queda, vandalismo ou abandono de objetos. Os métodos descritos a seguir avaliam vídeos em busca desse tipo de evento, voltado para detecção de comportamento anormal em cenários onde não há multidão.

A pesquisa apresentada por Perš et al. (2010) busca anomalias nas ações de uma pessoa em frente a um caixa de auto-atendimento. O método utiliza Histogramas de Fluxo Óptico para representar os movimentos na cena. Histogramas de Fluxo Óptico é uma forma de utilizar o FO para gerar um histograma de movimento para cada região da cena, baseado no acúmulo do fluxo. A técnica desenvolvida atua como um rastreador que identifica a região com maior variação de fluxo e cria blocos menores para avaliar as ações. A Figura 5 (a) exibe vetores do FO para um quadro do vídeo, Figura 5 (b) mostra o quadro do vídeo particionado em áreas menores e a Figura 5 (c) apresenta os níveis de histograma para cada área representados pela intensidade das marcações em vermelho.

Figura 5 – (a) FO para um quadro da cena, (b) quadro dividido em subáreas e (c) Histogramas de Fluxo Óptico para o mesmo quadro, representados pelas marcações em vermelho.



Fonte: Adaptado de Perš et al. (2010).

Os movimentos, descritos como um conjunto de símbolos, são classificados com base em sequências de símbolos pré-rotuladas. O método aprende um conjunto de ações para a cena e memoriza sequências de símbolos. Novas sequências são classificadas por vizinhança com a distância de Levenshtein[1]. Uma desvantagem do método apresentado por Perš et

[1] Algoritmo utilizado para calcular a similaridade entre duas cadeias de caracteres.

al. (2010) é a inconsistência temporal dos dados que resulta em falha de correspondência das sequências de símbolos. Esta falha ocorre devido à variação de tempo e ordem de execução das ações, que pode mudar de pessoa para pessoa.

De forma similar, a postura de uma pessoa é avaliada no trabalho desenvolvido por Chunli e Kejun (2010), que visa classificar o comportamento de uma pessoa como correr, pular ou caminhar. Essa pesquisa utiliza métodos baseados na aparência para gerar os modelos que descrevem cada comportamento. A classificação é feita pela proximidade das projeções geradas por extensões do PCA. As projeções que compõem o modelo não são atualizadas para cada imagem, como de costume. Os blocos de dados avaliados são construídos pelo acúmulo de energia de imagens sequenciais, abordagem chamada de Imagem da Energia de Caminhada. O modelo de descrição de movimento é passado como entrada para os métodos PCA, 2D-PCA e a extensão bidirecional bidimensional do PCA ($2D^2$-PCA).

Nos resultados dos testes apresentados por Chunli e Kejun (2010), o tempo e a memória utilizados pelo $2D^2$-PCA são menores em relação aos métodos PCA convencional e 2D-PCA. Quanto à qualidade das classificações, os resultados para o $2D^2$-PCA superam os demais métodos apenas com o aumento no números de amostras para treino. O trabalho mostra que combinar imagens vizinhas em uma sequência de vídeo pode auxiliar no reconhecimento de um comportamento. No entanto, uma desvantagem apresentada é a dependência do modelo em relação ao enquadramento das pessoas, pois o descritor baseado no acúmulo da energia deve representar o movimento de todo o corpo.

A Tabela 2 exibe os métodos citados nesta seção e trata algumas de suas características.

Tabela 2 – Métodos descritos que avaliam cenas sem multidão, comparados quanto ao tipo de comportamento, abordagem (global/local) e classificador utilizado.

Autores	Tipo de comportamento identificado	Avaliação global/local	Classificador
Perš et al. (2010)	Comportamento Anormal	Global e Local	kNN (dist. de Levenshtein)
Chunli e Kejun (2010)	Gestos Corporais	Local	$2D^2$–PCA

Fonte: O autor(2016).

3.3 Detecção de comportamento anômalo em vídeos com multidão

Analisar vídeos de segurança em cenários com multidão é uma atividade complexa, pois o número de pessoas na cena nem sempre é conhecido e pode não estar relacionado com o tipo de comportamento. Além disso, a qualidade dos vídeos de SVBVs, normalmente, é baixa, no que diz respeito ao nível de ruído na cena (ZIN et al., 2011). Os trabalhos a seguir apresentam propostas para avaliar cenas de vigilância em busca de comportamento anormal de multidão.

O uso de técnicas de aprendizagem de máquina que permitam classificar comportamentos é essencial para tarefas que envolvam sistemas de vigilância. No entanto, métodos baseados no FO, por exemplo, podem apresentar problemas quando se tenta cruzar informações de espaço e tempo. Originalmente, o FO gera modelos referentes à direção e velocidade em que o movimento na cena é realizado. Assim, para armazenar dados sem perder a consistência temporal do vídeo, Kaneko et al. (2014) propõem o uso de Campos Aleatórios Condicionais, modelos probabilísticos para rotular dados sequenciais. O método, capaz de identificar atividades coletivas, combina reconhecedores locais de atividade para a mesma cena.

A relação entre as ações isoladas é feita ao avaliar parâmetros de comportamento de escopo local, como: posição, tamanho, movimento e tempo de ocorrência. Os vários comportamentos são passados a um classificador SVM de múltiplas classes e cada pessoa é identificada como um vetor. As atividades são reconhecidas a nível de escopo global ao cruzar esses vetores com os parâmetros de comportamento. Uma desvantagem do método proposto por Kaneko et al. (2014) é sua fase de aprendizagem ser fortemente relacionada com a cena. Além disso, o método busca relacionar as pessoas após identificá-las na imagem, o que pode tornar o método inviável, quanto ao tempo de computação, em cenários com dezenas ou centenas de pessoas.

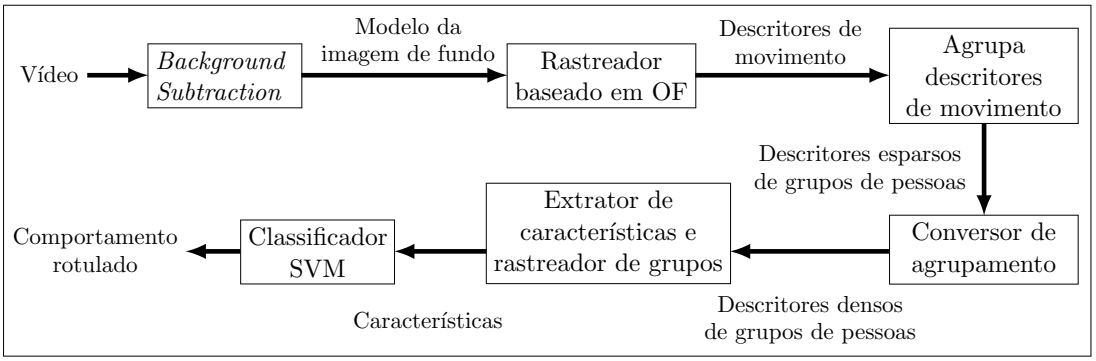
De forma similar, Chong et al. (2014) combinam a análise global e local para identificar trajetórias anormais em uma cena de multidão. Eles utilizam Processo Hierárquico de Dirichlet (do inglês, *Hierarchical Dirichlet Process* – HDP), modelo que descreve a distribuição probabilística dos dados e possibilita agrupar as variáveis, para modelar trajetórias na cena. Chong et al. (2014) propõem o uso de HMM para auxiliar a análise temporal – escopo global – do modelo, uma vez que o HDP possibilita avaliar o comportamento apenas no quesito espacial da cena, ou seja, análise de escopo local. Uma desvantagem do método é que a velocidade ou forma como as pessoas de movem não é levada em consideração, apenas o caminho que elas percorreram. Além disso, o método deve ser capaz de rastrear cada pessoa na cena para derivar um tipo de comportamento com base em sua trajetória.

Assim como na abordagem apresentada por Kaneko et al. (2014), o método proposto por Hassner, Itcher e Kliper-Gross (2012) utiliza SVM para classificar comportamentos, mas passa como características da cena conjuntos de descritores de fluxo gerados pelo FO. O trabalho visa identificar apenas dois tipos de comportamento: violento e não-violento. As cenas são rotuladas com base na magnitude e no sentido dos vetores de fluxo. O vídeo é cortado em pequenos trechos e assume-se que cada trecho contém somente um comportamento. Todos os trechos são divididos em células, para que histogramas de descrição da cena sejam criados. O método faz uma análise local e combina os histogramas de todo o trecho com os descritores de fluxo óptico para criar um conjunto do tipo *bag of visual features* que descreve o comportamento desse segmento do vídeo. Este conjunto de

descritores para cada corte, chamado de Fluxo Violento, é avaliado por um classificador SVM tanto para aprender o comportamento quanto para classificar novos trechos. Essa técnica teve bons resultados nos experimentos, mas os autores descrevem uma série de restrições quanto ao tipo de cena, qualidade dos vídeos, características dos descritores de Fluxo Violento e fase de treino do classificador SVM. Essas restrições podem tornar o método inviável em ambientes reais, por exemplo, em um SVBV.

O trabalho apresentado por Sindhuja, Srinivasagan e Kalaiselvi (2014) utiliza um classificador SVM multiclasse para reconhecer comportamentos em multidões. No entanto, o FO é implementado como um rastreador após aplicar *Background Subtraction*. Os vetores que descrevem o fluxo são utilizados para identificar agrupamentos de pessoas na cena. A Figura 6 exibe a arquitetura completa desse método e ilustra a sequência de execução dos módulos que a compõe. Características de cada agrupamento, como centro e orientação na cena, são extraídas e passadas para o classificador SVM multiclasse. No treino, são classificadas ações como: andar, correr, união e divisão de agrupamentos. As taxas de classificação obtidas nos experimentos foram elevadas, mas o conjunto de cenas que compõe a base utilizada nos experimentos é formado por quatro enquadramentos distintos da mesma ação. Esta característica da base de dados pode ter levado a um problema de sobreajuste – *overfitting*. Outra desvantagem do método é o fato de realizar todas as fases de sua arquitetura mesmo que a cena apresente apenas a imagem de fundo.

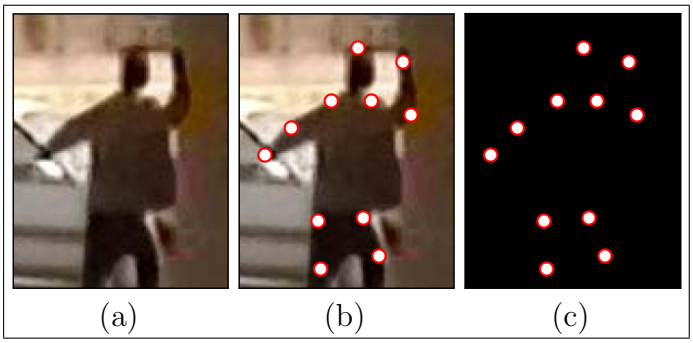
Figura 6 – Arquitetura proposta por Sindhuja, Srinivasagan e Kalaiselvi (2014).



Fonte: Adaptado de Sindhuja, Srinivasagan e Kalaiselvi (2014).

A pesquisa apresentada por Cohen et al. (2008) propõe uma arquitetura para avaliar sistemas de segurança e apresenta uma ferramenta para realizar experimentos e validar a arquitetura proposta. O trabalho define algumas das principais características que um sistema deve apresentar, tais como: (1) disparar sinais de detecção da anomalia, (2) exibir marcações que auxiliem os operadores humanos e (3) armazenar os vídeos em um servidor de dados que suporte mídias, para que avaliações futuras possam ser realizadas. Nos experimentos apresentados foram utilizados vídeos com vários comportamentos normais e anormais. Os vídeos foram exibidos a um grupo de operadores humanos junto com marcações do tipo *Point-light Walker* (PLW). As marcações PLW descrevem o movimento dos membros de uma pessoa por meio de pontos marcados sobre suas articulações, como mostra a Figura 7.

Figura 7 – (a) imagem com uma pessoa suspeita, (b) imagem e marcações *Point-light Walker* sobrepostas e (c) marcações *Point-light Walker* sem imagem de fundo.



Fonte: Adaptado de Cohen et al. (2008).

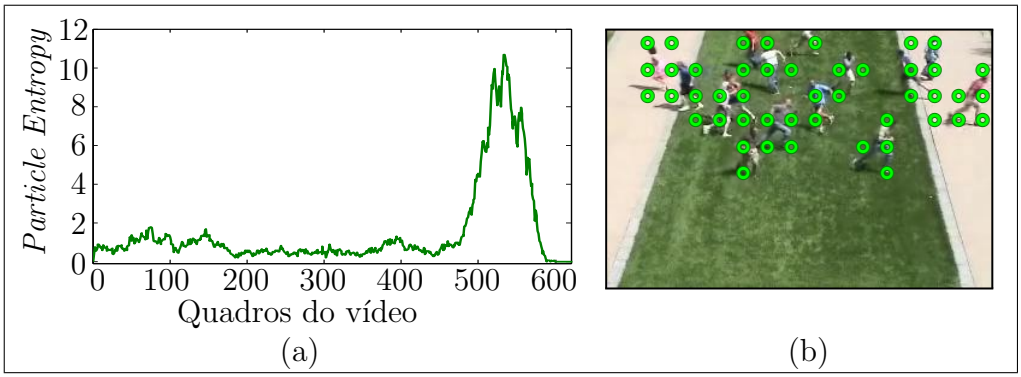
Ao visualizar os vídeos e as marcações, o operador humano poderia auxiliar o sistema inteligente com a inserção de novas marcações, criar áreas de interesse e rótulos para a cena. A arquitetura apresentada por Cohen et al. (2008) pode combinar parâmetros gerados de forma automática e manual para identificar um comportamento. O módulo responsável por classificar os comportamentos, baseado na Teoria de Dempster–Shafer, não foi avaliado. O plano de fundo é identificado com o método *Codebook Model Foreground-Background Segmentation* e as pessoas encontradas na cena são descritas por modelos que marcam separadamente tronco e membros.

O método apresenta ainda módulos para identificar objetos abandonados e gravar as trajetórias de pessoas rastreadas. Ainda que se trate de um trabalho inacabado, a completude do arcabouço apresentado por Cohen et al. (2008) e a preocupação quanto ao uso coerente de tecnologias inteligentes em sistemas reais de vigilância baseados em vídeo serviram de inspiração para a realização desta pesquisa e direcionaram o desenvolvimento do presente trabalho em aspectos como: respostas do sistema em tempo real, marcações visuais para auxiliar o operador humano e modularizarão da arquitetura para proporcionar ao método maior flexibilidade e escalabilidade.

Outra proposta desenvolvida para avaliar cenários de multidão em sistemas de vigilância é apresentada por Gu, Cui e Zhu (2014). A pesquisa combina descritores de FO e aprendizagem baseada em características em uma abordagem que utiliza o FO para derivar os movimentos em áreas menores, chamadas de partículas. A Figura 8 (a) mostra os níveis de movimento calculados para uma sequência de vídeo. A divisão das áreas simula uma grade, como ilustra a Figura 8 (b), e a quantidade de áreas varia conforme o tipo de cena. Caso a velocidade média do fluxo óptico de uma partícula seja inferior à um limiar, a área é tratada como plano de fundo e é eliminada das demais etapas do processo de reconhecimento de comportamento.

Em seguida, o método faz a contagem do número de partículas restante para gerar um índice de movimento chamado *Particle Entropy* (PE). A velocidade das partículas e o PE são as características do quadro atual do vídeo que descrevem o movimento e a dispersão

Figura 8 – (a) Níveis de movimento (*Particle Entropy*) para uma sequência de vídeo, (b) centros das partículas para um dos quadros do vídeo, marcados em verde.



Fonte: Adaptado de Gu, Cui e Zhu (2014).

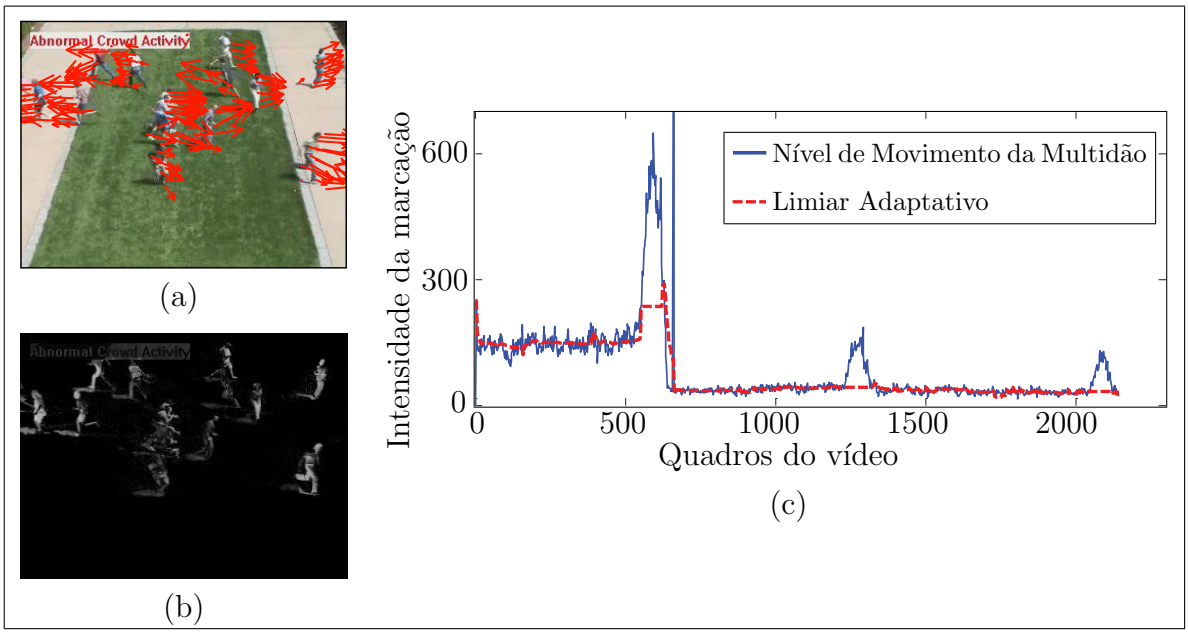
da multidão. Para reconhecer os comportamentos o método utiliza 300 valores de PE gerados no início do vídeo para criar um Modelo de Mistura Gaussiana (do inglês, *Gaussian Mixture Model* – GMM), modelo probabilístico utilizado para identificar subconjuntos no domínio das variáveis observadas. O GMM construído é utilizado para representar o comportamento normal presente na cena. A abordagem utiliza o início de cada vídeo da base de dados, que contém comportamento normal, para treinar o classificador GMM. Os valores que descrevem a dispersão e o movimento nos quadros subsequentes são rotulados com base no GMM. Nos experimentos, o método apresentou alta precisão em sua classificação, no entanto, foi testado em uma base que apresenta anormalidade somente na forma de dispersão das pessoas com mudança súbita entre o comportamento normal e o anormal. Além disso, um classificador GMM é treinado para cada vídeo, o que torna a abordagem altamente dependente dos modelos gerados para a cena corrente.

O trabalho apresentado por Liu, Li e Jia (2014) propõe uma abordagem baseada no FO, assim como Gu, Cui e Zhu (2014), mas que dispensa aprendizagem. O método apresenta etapas de pré-processamento, extração de características, pós-processamento e classificação. Para a etapa de pré-processamento, é feita a extração do plano de fundo baseada no *Background Subtraction*. A máscara gerada pelo extrator do plano de fundo descreve quais regiões na cena tiveram movimentos que ultrapassam um limiar. Em seguida, na etapa de extração de características, um método do tipo *Harris Corner Detector* é utilizado para selecionar pontos de interesse na máscara de movimentos. Assim, não são detectados pontos no plano de fundo. Os cantos são apresentados a um método esparso do FO com áreas de interesse para derivar o movimento na cena. A abordagem esparsa reduz o custo do FO, mas gera perda de continuidade no quesito temporal, uma vez que o detector de cantos não retorna pontos de interesse que tenham relação com os pontos detectados nos quadros vizinhos.

Os dados gerados pelo FO são convertidos em uma medida de movimento entre os quadros, chamada *Crowd Motion Intensity (CMI)*. Um método de limiar adaptativo é

proposto para avaliar o CMI e identificar o momento onde ocorre a mudança de comportamento. A Figura 9 (a) apresenta o fluxo esparso que representa o movimento entre dois quadros do vídeo. A Figura 9 (b) mostra a máscara gerada pelo *Background Subtraction* que foi passada para o detector de cantos. Os níveis CMI derivados para uma sequência de vídeo podem ser vistos em azul na Figura 9 (d) e o limiar adaptativo, em vermelho, também na Figura 9 (d). Uma desvantagem do método é sensibilidade a movimentos que não reflitam as ações das pessoas presentes na imagem. Isso ocorre dada a alta dependência que o método apresenta do modelo de plano de fundo, que propaga o erro para as demais etapas do processo.

Figura 9 – (a) fluxo óptico esparso para o movimento entre dois quadros do vídeo, (b) máscara gerada pelo *Background Subtraction* para o mesmo quadro, (c) níveis de movimento e valores do limiar adaptativo para uma sequência de vídeo, nas cores azul e vermelho, respectivamente.



Fonte: Adaptado de Liu, Li e Jia (2014).

As bordagens apresentadas por Liu, Li e Jia (2014) e Gu, Cui e Zhu (2014) foram incluídas nos experimentos e serão utilizadas como *baselines*. Eles foram escolhidos por utilizarem FO para analisar cenas de multidão, assim como o método aqui proposto, e em seus experimentos ambos apresentaram alta precisão ao marcar o comportamento anormal na base de dados. Além disso, o método proposto por Liu, Li e Jia (2014) aplica uma heurística de limiar auto-adaptado, sem aprendizagem, enquanto que a abordagem apresentada por Gu, Cui e Zhu (2014) identifica o comportamento na cena ao usar um classificador GMM após uma fase de aprendizagem.

A Tabela 3 exibe os métodos citados nesta seção e compara algumas de suas características. Na Seção 5.2, os dois métodos são comparados com o método proposto nesta pesquisa e a Seção 5.3 discute os resultados dos experimentos realizados.

Tabela 3 – Métodos discutidos que avaliam com multidão, comparados quanto ao tipo de comportamento, abordagem (global/local) e classificador utilizado.

Autores	Tipo de comportamento identificado	Avaliação global/local	Classificador
Cohen et al. (2008)	Comportamento Anormal	Global e Local	Dempster-Shafer
Hassner, Itcher e Kliper-Gross (2012)	Comportamento Anormal	Global	SVM
Sindhuja, Srinivasagan e Kalaiselvi (2014)	Comportamento Anormal	Global	SVM multi-classe
Kaneko et al. (2014)	Atividades coletivas	Global e Local	SVM multi-classe
Chong et al. (2014)	Comportamento Anormal	Global e Local	HDP
Gu, Cui e Zhu (2014)	Comportamento Anormal	Global	GMM
Liu, Li e Jia (2014)	Comportamento Anormal	Global	–

Fonte: O autor(2016).

3.4 Considerações Finais

Ao avaliar os trabalhos existentes na literatura foi observado que os métodos normalmente combinam técnicas diferentes, mas sem ponderar seu uso com base no conteúdo de vídeo recebido para análise. Por exemplo, o método proposto por Sindhuja, Srinivasagan e Kalaiselvi (2014) utiliza FO em uma de suas fases para gerar os vetores que descrevem os movimentos da cena. No entanto, mesmo quando a cena não apresenta qualquer movimento, todas as demais etapas são executadas. Este tipo de abordagem ocorre tanto em métodos baseados na aparência quanto nos métodos baseados em caraterísticas.

Assim, executar os mesmos procedimentos para todo o conteúdo recebido na entrada, seja de comportamento normal ou não, limita o uso dessas abordagens em aplicações reais com SVBVs de múltiplas câmeras. Um fator que agrava esse cenário é o uso de estratégias que aumentam a complexidade de computação conforme o número de pessoas na cena cresce, como pode ser visto nos trabalhos apresentados por Saini et al. (2012) e Zin et al. (2014). A Tabela 4 expõe algumas das principais características dos métodos descritos nesta seção.

A tabela mostra que somente cinco dos dezesseis métodos avaliam as cenas em escopo global e local. Um sistema inteligente deve marcar o momento onde uma anomalia ocorre. Essa tarefa pode ser alcançada por uma aplicação de escopo global. No entanto, gerar marcações na cena para auxiliar o operador humano requer que o método avalie a cena em escopo local. Caso o sistema opere somente como aplicação local, o reconhecimento do comportamento pode ser falho, como ocorre na pesquisa apresentada por Perš et al. (2010). Assim, um sistema inteligente capaz de operar como aplicação global e local, no mesmo vídeo, pode identificar um comportamento anormal através da análise de tempo e espaço da cena e gerar marcações mais confiáveis.

Na prática, SVBVs possuem múltiplas câmeras e fazem a aquisição de vários vídeos ao mesmo tempo. Executar métodos de forma igualitária a todos os vídeos pode acarretar um elevado custo de computação. Assim, a abordagem apresentada neste trabalho faz análises

Tabela 4 – Compilação dos métodos descritos, comparados quanto ao tipo de características extraídas, forma de representação do comportamento, suporte a cenas de multidão, tipo de abordagem (global/local) e classificador utilizado.

Autores	Características Extraídas	Representação do Comportamento	Suporta Multidões	Aplicação Local/Global	Classificador
(ALI; SHAH, 2007)	Sentido e velocidade (FO)	Segmentos (Op. de Divergência)	Sim	Global	–
(COHEN et al., 2008)	Silhueta / Contorno, Trajetória	Postura, posição e velocidade	Não	Global e Local	Dempster-Shafer
(CHUNLI; KEJUN, 2010)	Silhueta / Bloco de Pixels	Imagem da Energia de Caminhada	Não	Local	$2D^2$–PCA
(PERŠ et al., 2010)	Histogramas de FO	Posição e velocidade (FO)	Não	Global e Local	kNN (dist. de Levenshtein)
(BROX; MALIK, 2011)	Sentido e velocidade (FO)	LDOF	Sim	Global	–
(JIANG et al., 2011)	Posição e Sentido	Trajetória	Sim	Global	MOM e Vizinhos mais próximos
(HASSNER et al., 2012	Histogramas de FO	Sentido e velocidade (FO)	Sim	Global	SVM
(JIANG et al., 2013)	Silhueta / Contorno, Movimento	Postura e Posição	Não	Local	HMM e SVM
(CHONG et al., 2014)	Áreas de interesse	Localização e Histogramas orientados à Velocidade	Sim	Global e Local	HDP
(SEO; KIM, 2014)	Componentes principais	–	Não	Global	$2D^2$–PCA
(SINDHUJA et al., 2014)	FO	Sentido e velocidade (FO)	Sim	Global	SVM multi-classe
(KANEKO et al., 2014)	Posição, tamanho, movimento e sequência de tempo	Campos Aleatórios Condicionais	Sim	Global e Local	SVM multi-classe
(LIU; LI; JIA, 2014)	FO	Escala do FO	Sim	Global	–
(GU; CUI; ZHU, 2014)	FO	*Particle Entropy*	Sim	Global	GMM
(ZIN et al., 2014)	Movimento / Trajetória	Movimento e posição	Não	Global e Local	Cadeia de Markov
Método proposto	Descritor da aparência da cena e FO	Estrutura da cena e intensidade dos movimentos	Sim	Global e Local	–

Fonte: O autor (2016).

globais e locais em fases independentes. Na análise global, cenas que possam conter comportamento anormal são selecionadas. O método faz a análise local apenas em cenas que apresentam comportamento suspeito. Portanto, o método proposto difere das arquiteturas convencionais para reconhecimento de comportamento em vídeo, pois, ao avaliar um comportamento na cena, não executa todos os procedimentos que compõem a arquitetura, mas apenas os procedimentos necessários ao contexto. Cohen et al. (2008) apresentam uma arquitetura que seleciona os vídeos com comportamento suspeito. Para auxiliar o operador humano, o método exibe marcações de objetos ou pessoas que geraram o alarme do comportamento anormal. No entanto, durante o processo de reconhecer o comportamento, não é apresentada uma estratégia voltada para a redução do custo de computação ao avaliar os vídeos.

O próximo capítulo apresenta o método proposto neste trabalho e descreve em detalhes o funcionamento de cada um de seus módulos.

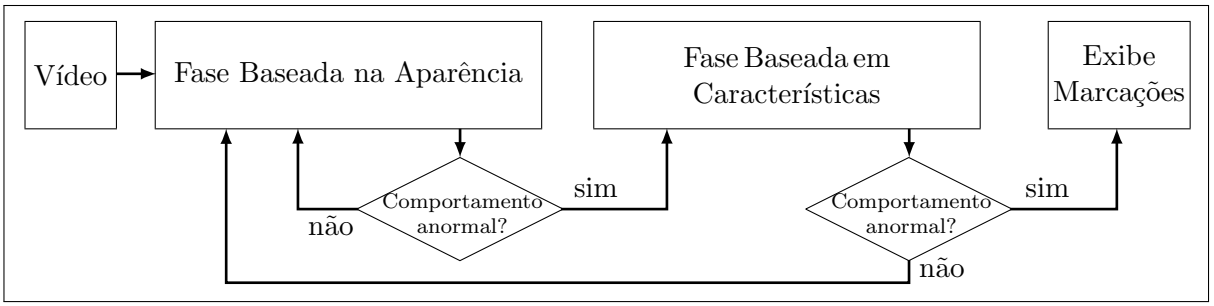
4 Método Proposto

Este capítulo apresenta a estrutura e o funcionamento do método proposto. A Seção 4.1 exibe um diagrama com a estrutura do método proposto, descreve o funcionamento de cada módulo que o compõe e como eles se inter-relacionam. A Seção 4.2 discute as características do método proposto e como ele se comporta em relação aos métodos encontrados na literatura.

4.1 Arquitetura do Método

A Figura 10 exibe a arquitetura proposta, que é composta por duas fases: a baseada na aparência e a baseada em características. Cada uma das fases pode ser formada por diferentes abordagens capazes de avaliar o vídeo na entrada e, caso necessário, gerar marcações na imagem para auxiliar o operador humano a visualizar a anomalia detectada.

Figura 10 – Representação genérica da arquitetura proposta.

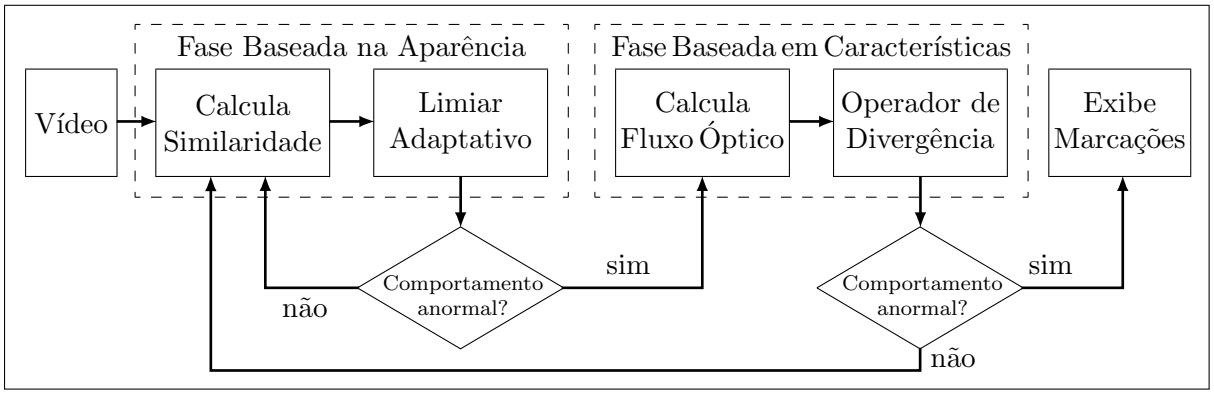


Fonte: O autor (2016).

Na fase baseada na aparência foram investigados seis métodos para avaliar a estrutura da cena sem realizar pré-processamento ou extrair características. Para essa fase, foi proposta uma heurística de limiar adaptativo responsável por ativar a segunda fase da arquitetura quando um comportamento anormal é detectado. A Fase baseada em características é composta pelo FO e por um operador de divergência. A Figura 11 apresenta a estrutura do método proposto dividida em módulos. Os componentes que compõem a arquitetura podem ser substituídos por alternativas que se adequem a um cenário ou problema específico, como por exemplo, o uso de um classificador para substituir a heurística de limiar adaptativo, na fase baseada na aparência, ou para classificar os descritores do fluxo óptico na fase baseada em características. Os elementos que compõem o método proposto são descritos a seguir.

O módulo de entrada do método recebe dados de vídeo com cenas obtidas por um sistema de vigilância. Normalmente, as cenas são de câmeras fixas posicionadas em locais estratégicos, como: corredores, passarelas, faixas de pedestre e nas principais portas de

Figura 11 – Método proposto baseado na arquitetura apresentada, dividido em duas fases: baseada na aparência e baseada em características.



Fonte: O autor (2016).

acesso à estabelecimentos. Dessa forma, a abordagem proposta foi modelada para cenários com câmeras estáticas. Em seguida, o método inicia a verificação do vídeo com base na aparência, fase apresentada com mais detalhes na Seção 4.1.1. Essa fase é responsável por avaliar o vídeo em busca de comportamentos suspeitos e, caso necessário, iniciar o segundo módulo da arquitetura, a fase baseada em características. A fase identifica possíveis falsos positivos gerados pela primeira ou confirma a decisão do módulo anterior ao exibir marcações do comportamento suspeito para auxiliar o operador humano. A fase baseada em características é descrita em detalhes na Seção 4.1.2.

4.1.1 Fase Baseada na Aparência

Esta fase avalia a cena como uma aplicação global e pode utilizar qualquer método baseado na aparência, como por exemplo, um dos métodos descritos na Seção 2.2.2 (SSIM, MSM, PCA, 2D-PCA, IPCA ou IMED), para derivar o nível de variação na estrutura da cena. Esse nível representa a intensidade das mudanças que ocorreram em um intervalo de tempo do vídeo, entre o quadro atual e o seguinte. Essa abordagem é baseada em trabalhos que classificam imagens ao avaliar sua aparência, como Kim et al. (2013) e Kim e Park (2012). Nessa abordagem, os elementos visuais que compõem a cena são representados pelos itens que descrevem a estrutura da imagem como: contornos, linhas, e pontos. Utilizar métodos baseados na aparência possibilita estimar uma medida que descreve o nível de similaridade entre duas ou mais imagens com base na representação de sua estrutura.

No início da fase, duas imagens são entregues ao método que irá gerar o nível de similaridade. Os métodos do subespaço e as abordagens SSIM e IMED não irão avaliar informações de cor. Caso as imagens sejam coloridas, elas são convertidas em matrizes que descrevem apenas níveis de intensidade. Seja I uma imagem colorida de dimensões $n \times m \times 3$, onde a última dimensão descreve o número de canais de cores do padrão RGB. A imagem I pode ser representada pelas matrizes I_R, I_G e I_B, de dimensão $n \times m$ com

valores inteiros entre 0 e 255, para os níveis de coloração dos canais vermelho, verde e azul, respectivamente. A conversão de I para uma imagem em níveis de intensidade I_{gray} pode ser feita pela fórmula:

$$I_{gray} = (0.299 \times I_R) + (0.587 \times I_G) + (0.114 \times I_B). \tag{4.1}$$

O número de quadros que esta fase utiliza para derivar os níveis de similaridade pode ser ajustado por um parâmetro ω. Quando $\omega = 1$, o método deriva a similaridade entre as imagens I_t e $I_{(t+1)}$, com t para o tempo do vídeo. Caso $\omega > 1$, o método cria duas imagens médias, I_1 e I_2, onde:

$$I_1 = \frac{1}{\omega}\sum_{t}^{\omega} I_t \qquad \text{e} \qquad I_2 = \frac{1}{\omega}\sum_{(t+\omega)}^{2\ \omega} I_t. \tag{4.2}$$

A mudança do parâmetro ω ameniza problemas causados por ruído na cena e normaliza os níveis de similaridade gerados, principalmente para vídeos com baixas taxas de quadros por segundo. Nos experimentos realizados neste trabalho, o controle do parâmetro ω com valores entre 3 e 5 melhorou consideravelmente a qualidade dos resultados em vídeos com taxa de 15 quadros por segundo ou menor. A suavização ocorre, pois, ao fazer a média das imagens, regiões de textura são uniformizadas enquanto o número de contornos na cena aumenta. Assim, modificar o valor de ω permite controlar a sensibilidade do método quanto ao tipo de vídeo gerado pelo SVBV. Nos experimentos, foi observado que, em vídeos com taxa de quadros por segundo superior a 15, o método apresenta resultados melhores para $\omega = 2$.

Em seguida, a fase baseada na aparência inicia uma heurística de limiar auto-adaptado que irá avaliar os valores de similaridade, descrita em detalhes a seguir. Essa abordagem foi escolhida pois dispensa fase de treino com dados acumulados em memória, o que possibilita a aplicação do método em ambientes reais. Os valores no início do vídeo são utilizados como descritores de comportamento normal da cena. Assim, os dados gerados na própria sequência de vídeo são utilizados para ajustar os níveis do limiar auto-adaptado. Como descrevem Bouzegza e Elarbi-Boudihir (2013), heurísticas de limiar auto-adaptado podem ser comparadas com técnicas de aprendizagem não supervisionada. Uma vantagem ao utilizar essa abordagem é não ter que ajustar parâmetros para cada cena. Liu et al. (2013) afirmam que esta estratégia é semelhante ao funcionamento do cérebro humano ao trabalhar com memória curta e é adotada em várias pesquisas por apresentar bons resultados em aplicações reais.

A heurística apresentada a seguir foi desenvolvida com base nos fundamentos aplicados em algoritmos de detecção de *drift* para dados em *streaming*. Métodos como DDM (*Drift Detection Method*) e EDDM (*Early Drift Detection Method*) se baseiam na distribuição dos dados lidos na entrada para determinar um nível de advertência e um nível de detecção

de *drift* (GAMA et al., 2004; BAENA-GARCIA et al., 2006). Esses níveis atuam com limiares auto-ajustados que descrevem se houve mudança na distribuição dos dados. A heurística desenvolvida para esta pesquisa dispensa o uso de algoritmos de aprendizagem de máquina, elemento presente nesses métodos de detecção de *drift*.

A heurística mantém um conjunto de valores em memória – *buffer* – para armazenar os últimos índices de similaridade gerados. O tamanho do *buffer* pode ser ajustado pelo parâmetro β com valor 1 ou maior. Em nossos experimentos, o método apresentou melhores resultados ao utilizar valores para β entre 3 e 15. O método acumula ainda a média e a variância dos índices que passam pelo *buffer*. Ambos são obtidos de forma incremental, para que os dados removidos do *buffer* não tenham que ser armazenados em memória. A média incremental $\bar{D}_t$ para os valores de similaridade no momento t do vídeo, pode ser calculada pela fórmula:

$$\bar{D}_t = t^{-1}[D_t + (t-1)\ \bar{D}_{(t-1)}] \tag{4.3}$$

onde D_t é o último valor de similaridade gerado no tempo t. O desvio padrão s_t^2 dos valores de similaridade lidos até o tempo t pode ser obtido de forma incremental pela equação:

$$s_t^2 = \frac{t-2}{t-1} s_{(t-1)}^2 + \frac{1}{t}\ (D_t - \bar{D}_{(t-1)})^2. \tag{4.4}$$

Após obter a média e o desvio padrão, o método avalia os novos valores de similaridade carregados para o *buffer* na busca por níveis acima do limiar auto-adaptado $\alpha = (\ D_t + s_t^2\ )$. O valor de α é atualizado a cada β valores de similaridade lidos, caso o *buffer* não tenha apresentado nenhum valor maior que α neste intervalo. Ao iniciar a heurística de limiar adaptativo, $\bar{D}_1$ e s_1^2 recebem o primeiro nível de similaridade. Neste momento, a heurística está em fase de adaptação e os dados lidos são tratados como comportamento normal. Durante os experimentos, foi observado que a quantidade de níveis de similaridade necessária para a fase de adaptação varia entre 2 e 3 vezes o valor de β, conforme os valores de α, β, e da taxa de quadros por segundo do vídeo. Dessa forma, para um *buffer* de tamanho 4, são necessários de 8 a 12 quadros do vídeo para que o limiar seja ajustado à cena.

Quando a etapa de adaptação da heurística é finalizada, o método inicia a análise dos níveis de similaridade. Valores superiores a α marcam a cena como comportamento suspeito e o método passa para a etapa baseada em características, descrita em detalhes na Seção 4.1.2. Para os casos onde os níveis de similaridade são iguais ou menores que α, a cena é tratada como comportamento normal e o método dá continuidade à análise do vídeo baseada na aparência.

4.1.2 Fase Baseada em Características

Quando a fase baseada na aparência identifica um trecho de vídeo com comportamento suspeito, a fase de reconhecimento baseada em características é inicializada. Essa fase verifica se a marcação do comportamento suspeito corresponde a um falso positivo. Caso o comportamento anormal seja confirmado, o método verifica o tipo de comportamento e a região da cena onde a ação é observada. Caso contrário, a fase baseada em características é finalizada e o método dá continuidade à verificação do vídeo pela fase baseada na aparência.

A fase baseada em características utiliza o Fluxo Óptico (FO) para derivar o conjunto de vetores que estimam o fluxo na cena. Os descritores do FO são utilizados como características que descrevem o comportamento no escopo global e local da cena. Assim como no método apresentado por Hassner, Itcher e Kliper-Gross (2012), o comprimento dos vetores que descrevem o fluxo é utilizado como índice de intensidade do comportamento presente no vídeo, calculado por:

$$c = \sqrt{x^2 + y^2} \tag{4.5}$$

onde x e y são os componentes que descrevem um vetor $\vec{v}$ no espaço 2D e c é o comprimento obtido para o vetor $\vec{v}$. Vetores com comprimento inferior a um limiar são tratados como ruído e são removidos.

Como descrito na Seção 3.1, o trabalho apresentado por Ali e Shah (2007) trata a multidão como um conjunto de partículas que se desloca na cena. Os movimentos no vídeo são avaliados com base nos níveis de dispersão e aglomeração calculados por um operador de divergência e utiliza essas informações para segmentar trechos da cena com padrões distintos de deslocamento. De maneira semelhante ao trabalho apresentado por Ali e Shah (2007) este trabalho utiliza o FO na fase baseada em características como um descritor dos movimentos de partículas que fluem no espaço da imagem. Um operador de divergência é aplicado para identificar as regiões na imagem para onde as partículas convergem ou de onde elas divergem. No entanto, o método proposto não segmenta a cena e utiliza os níveis de convergência e divergência para identificar regiões de aglomeração e de dispersão das pessoas no vídeo.

O operador de divergência utilizado na fase baseada em características é descrito da seguinte forma: Seja o conjunto $\vec{\nu}$ de vetores do fluxo óptico, em coordenadas cartesianas, com dimensões $m \times n$ para uma imagem I com as mesmas dimensões. O operador de divergência $(\nabla \cdot \vec{\nu})$ para o conjunto $\vec{\nu}$ pode ser descrito pela fórmula:

$$\nabla \cdot \vec{\nu} = \sum_{i=1}^{m \times n} \frac{\partial \vec{\nu}_i}{\partial \delta_i} \tag{4.6}$$

onde δ é um conjunto com $m \times n$ coordenadas de um sistema cartesiano no espaço euclidiano que relacionam cada vetor de $\vec{\nu}$ a seu respectivo pixel da imagem I. Ao obter os

valores de divergência $\lambda = (\nabla \cdot \vec{\nu})$ para a cena suspeita, a fase baseada em características divide a imagem em k regiões menores, chamadas células.

Cada célula C_k é rotulada com um tipo de comportamento, conforme a equação:

$$C_k = \begin{cases} \text{aglomeração} & \text{se } \lambda_k > \phi \\ \text{dispersão} & \text{se } \lambda_k < \phi \\ \text{normal} & \text{caso contrário} \end{cases} \tag{4.7}$$

onde λ_k é o valor absoluto da média dos níveis de divergência na região correspondente à célula k. O valor de ϕ é um limiar para tratar valores próximos a zero, ou seja, níveis de divergência/convergência pequenos são marcados como comportamento normal. Dividir a cena em células para avaliar o comportamento na cena é uma abordagem utilizada em trabalhos como Krausz e Bauckhage (2012), Bertini, Bimbo e Seidenari (2012) e Gu, Cui e Zhu (2014). Os trabalhos utilizam também a estratégia para exibir marcações ao operador humano, como ilustra a Figura 12.

Figura 12 – Cena com marcação visual em vermelho das células onde foi identificado aumento da densidade de pessoas na cena.

Fonte: Adaptado de Krausz e Bauckhage (2012).

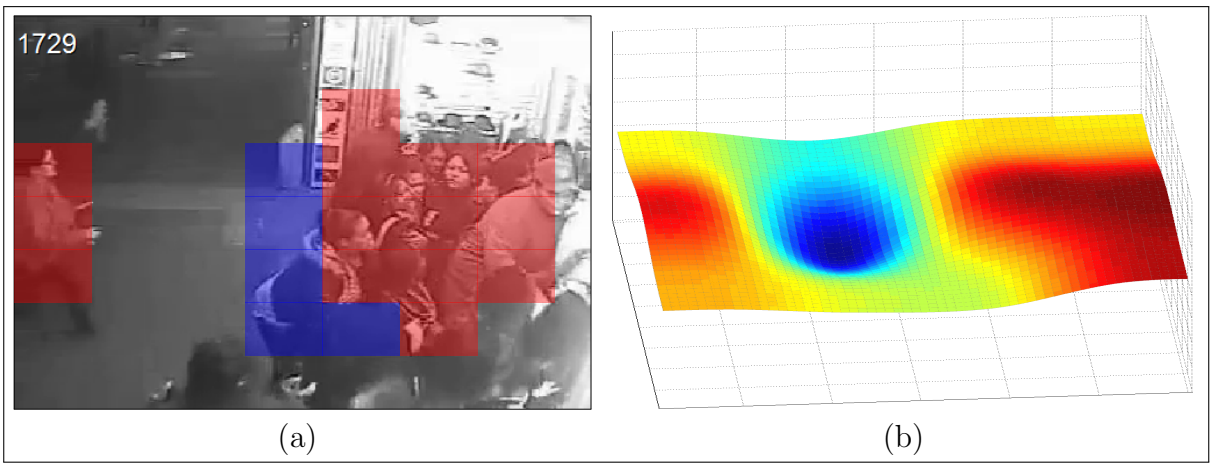
Em seguida, a fase baseada em características avalia a ocorrência de cada classe no conjunto C_k. O quadro do vídeo no tempo t, até então marcado como suspeito pelo método baseado na aparência, é rotulado como comportamento anormal caso o percentual de células de aglomeração e dispersão juntas em C_k seja superior a um parâmetro γ. Caso contrário, o método confirma um falso positivo da fase baseada na aparência ao marcar o quadro como normal. Com isso, o comportamento na abordagem local não é avaliado e a fase baseada em características é desativada.

O parâmetro γ é ajustado conforme o enquadramento das pessoas na cena, visto que, uma única pessoa pode ocupar grande área da imagem ou somente alguns pixels, conforme sua distância da câmera. O tamanho das células varia de acordo com a resolução do vídeo e o enquadramento das pessoas. Células pequenas podem acarretar em um processo com elevado custo computacional e em uma saída confusa de ser avaliada pelo operador humano. Por outro lado, células grandes podem resultar em uma descrição imprecisa, principalmente quando houver vários comportamentos na mesma cena. Assim, as células

são criadas de forma que elas ocupem uma região quadrada com área de 8% a 20% do tamanho da imagem exibida, valores definidos ao avaliar visualmente o tamanho das marcações durante os experimentos realizados neste trabalho.

Assim que o método confirma a existência de comportamento anormal na abordagem global, é iniciada a marcação no escopo local. De forma semelhante à Figura 12, a imagem I recebe marcações divididas em células. As células que compõem o conjunto C_k rotuladas como aglomeração são marcadas nas áreas correspondentes da imagem em vermelho. Células com rótulo de dispersão são marcadas em azul. As demais áreas da imagem permanecem inalteradas. A Figura 13 (a) mostra uma cena marcada como comportamento anormal pela fase baseada em características. A Figura 13 (b) exibe os níveis de divergência/convergência gerados pelo operador ∇ para o mesmo trecho do vídeo. O vídeo foi marcado com células de aglomeração e dispersão, nas cores vermelho e azul, respectivamente. Neste quadro do vídeo, as pessoas na cena se deslocam para a direita ao tentar entrar em uma loja.

Figura 13 – (a) cena com marcação das células de dispersão em azul e aglomeração, marcadas em vermelho e (b) níveis de divergência/convergência gerados pelo operador ∇ para o mesmo intervalo do vídeo.



Fonte: O autor (2016).

Os sinais gerados pelo método são mostrados para o operador humano sobre o vídeo e por meio de um relatório. No relatório são salvas as informações de tempo onde o comportamento anormal foi inicialmente identificado e quais os tipos de células foram marcadas com este comportamento. Quando nenhum comportamento anormal é detectado, a fase baseada em características é desativada e o sistema volta a operar com o método baseado na aparência.

4.2 Considerações Finais

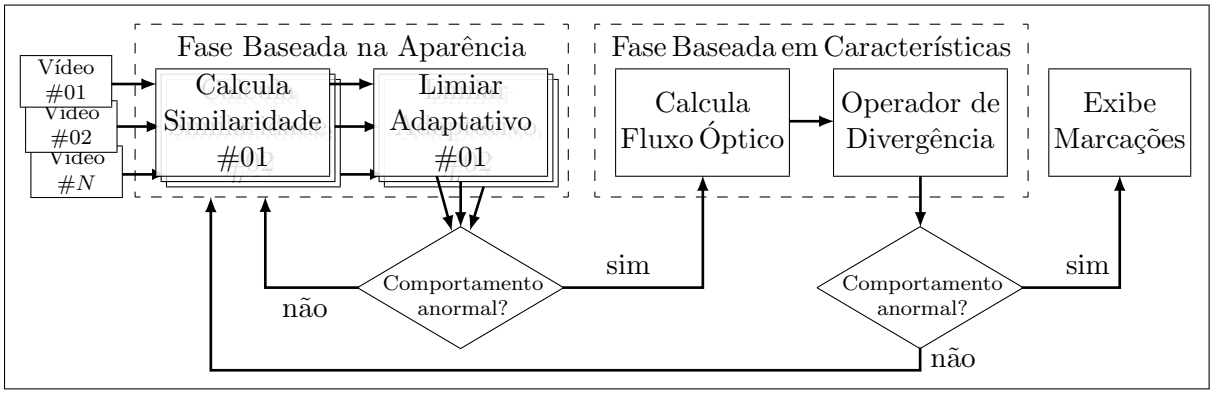
O método proposto faz uso de dois reconhecedores de comportamento, um baseado em características e outro baseado na aparência. Esta escolha visa utilizar cada tipo de abordagem de forma mais aderente ao problema em questão. Métodos baseados na aparência

apresentam grande robustez no que diz respeito à identificação de padrões em grupos de imagens. Um dos principais motivos está relacionado com sua capacidade em manter a consistência espacial dos dados. Os métodos baseados em características possibilitam que mais detalhes sejam investigados nos vídeos, tais como dividir comportamentos distintos em várias áreas e identificar o sentido do movimento na cena. No entanto, métodos que extraem características podem ser complexos quanto ao custo de computação e uso de memória. Por este motivo, a fase baseada em características do método proposto é aplicada apenas quando existe uma suspeita de comportamento anormal na cena, segundo a análise da fase baseada na aparência.

Esta estratégia assume que, em sistemas de segurança reais, grande parte do conteúdo de vídeo adquirido é referente a comportamento normal. Dessa forma, não se faz necessário aplicar todas as etapas do método de reconhecimento de comportamento para identificar ações anormais no vídeo que contém, por exemplo, apenas a imagem de fundo com algum ruído, movimentos da vegetação ou mudanças de iluminação na cena. Estratégias que visam a redução do custo de computação e de uso da memória possibilitam que SVBVs tenham múltiplas cenas avaliadas paralelamente. Este é um fator importante, uma vez que esse tipo de sistema, normalmente, faz a aquisição de vários vídeos ao mesmo tempo.

A Figura 14 mostra como a arquitetura do método pode ser adaptada para receber como entrada vários vídeos ao mesmo tempo de um sistema de vigilância que realiza a aquisição simultânea com várias câmeras. Assim, o método proposto pode ser aplicado em SVBVs para que múltiplos vídeos sejam avaliados. A arquitetura apresentada permite que sejam utilizadas outras estratégias de reconhecimento de comportamento baseadas em características ou na aparência, conforme peculiaridades do problema, da cena ou do sistema de vigilância em uso.

Figura 14 – Arquitetura do método proposto adaptada para um SVBV com múltiplas câmeras.



Fonte: O autor (2016).

A seguir, o Capítulo 5 apresenta os experimentos realizados com o método proposto, descreve as bases de dados investigadas e compara o método aqui apresentado com outros dois métodos presentes na literatura.

5 Experimentos e Resultados

Para avaliar a qualidade das marcações geradas pelo método proposto, os experimentos foram divididos em duas séries. A primeira avalia se os métodos baseados na aparência descritos na Seção 2.2.2 são capazes de detectar mudança de comportamento em vídeos como aplicações de escopo global. Esses métodos são responsáveis por gerar marcações na linha de tempo do vídeo onde um comportamento suspeito teve início. Os melhores métodos identificados nessa série de experimentos são testados na fase baseada na aparência da arquitetura proposta.

A segunda série de experimentos combina as duas fases que compõem a arquitetura e avalia o método proposto em dois aspectos: tempo de processamento e qualidade das marcações nas abordagens global e local. Nesta segunda série, a proposta desta pesquisa é comparada com outros dois trabalhos: Liu, Li e Jia (2014), baseado em heurística de limiar auto-adaptado e Gu, Cui e Zhu (2014), que utiliza um classificador GMM treinado com o início do vídeo.

Os experimentos foram realizados em 3 bases, duas públicas e uma criada para este trabalho. A base utilizada na primeira série de experimentos é descrita na Seção 5.1.1. As bases usadas na segunda série de experimentos são apresentadas na Seção 5.2.1. A seção a seguir explica a metodologia aplicada nos experimentos. A Seção 5.3 discute de forma breve, os conceitos abordados neste capítulo e compara os métodos tratados neste trabalho.

5.1 Comparação entre Métodos Baseados na Aparência

A primeira série de experimentos é voltada para a etapa de reconhecimento da mudança de comportamento baseado na aparência. O vídeo é avaliado por intervalos de quadros adjacentes. O comprimento desse intervalo é definido pelo parâmetro ω, como descrito na Seção 4.1.1. Uma medida de similaridade é gerada para esse intervalo e o conjunto de similaridades do vídeo descreve os valores de similaridade para cada trecho da gravação. A heurística de limiar auto-adaptado é aplicada ao conjunto de similaridades. Quadros marcados como suspeitos são comparados com os valores definidos por observação prévia – *ground truth*. A base de dados utilizada nos experimentos é descrita na seção a seguir.

5.1.1 Base de Dados Investigada

Na primeira série de experimentos foi utilizada a base *Unusual Crowd Activity*[1], da Universidade de Minnesota (UMN). A base apresenta cenas internas e externas onde pessoas

[1] Unusual Crowd Activity Dataset. Artifical Intelligence, Robotics and Vision Laboratory (AIRVL). University of Minnesota, Department of Computer Science and Engineering. 2006. Disponível em: <http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi>.

simulam um comportamento anormal. No início, o comportamento presente é normal, com pessoas que andam pela cena. Em seguida, as pessoas correm para fora da cena, trecho do vídeo marcado como comportamento anormal. A base UMN contém 11 tomadas, divididas em 3 ambientes. A Figura 15 mostra uma imagem de cada ambiente presente na base.

Figura 15 – Os três ambientes da base de dados UMN.

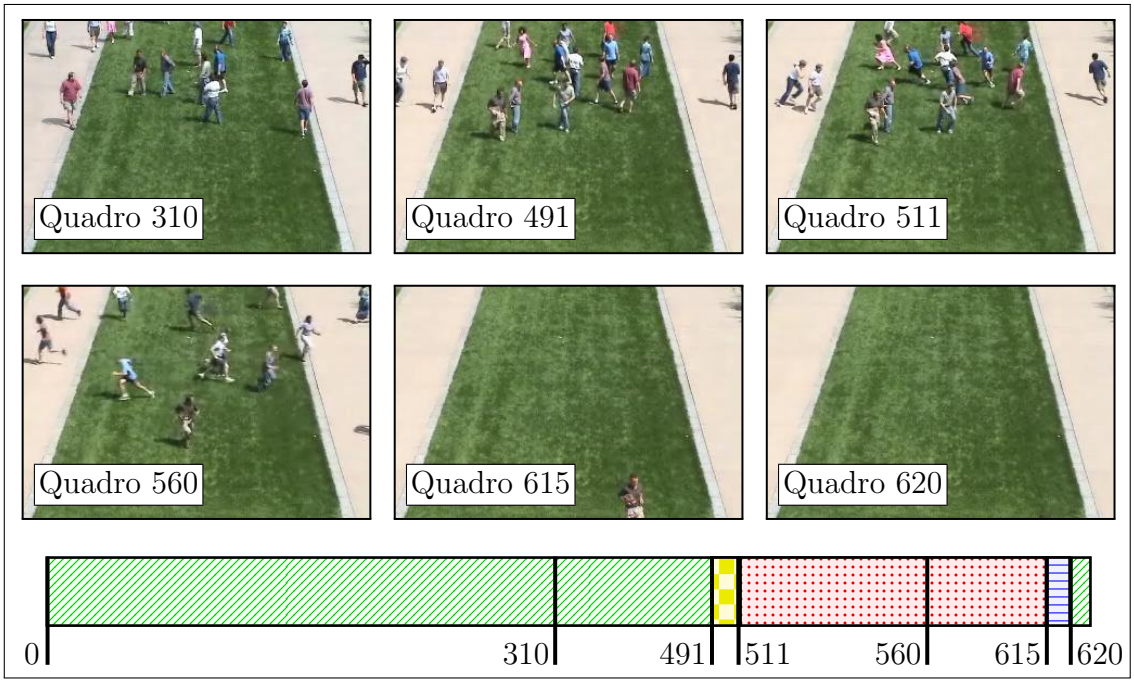
Fonte: Adaptado de *Unusual Crowd Activity Dataset*, AIRVL, Universidade de Minnesota (2006).

As marcações originais da base apresentam os quadros onde o comportamento anormal tem início. No entanto, como relata Chen e Huang (2011), visualmente a mudança de comportamento tem início antes do local marcado na base. Além disso, a mudança gradual de comportamento é passível de interpretações divergentes, o que dificulta medir a qualidade das marcações geradas por um sistema inteligente. Dessa forma, foi criado um novo conjunto de marcações para a base UMN através de uma enquete aplicada a 18 pessoas que não tinham conhecimento do funcionamento dos métodos tratados nesta pesquisa.

Os participantes da enquete foram instruídos a marcar o quadro do vídeo onde o início e o fim do comportamento anormal eram visíveis. Assim, os vídeos tiveram dois quadros marcados que descrevem o intervalo onde os participantes da enquete identificaram o comportamento anormal. No entanto, foi observada divergência nas marcações, o que nos dá dois intervalos de quadros, e não apenas dois quadros, o de início e o de fim do comportamento anormal. Com isso, nesta série de experimentos, a base UMN foi avaliada segundo as novas marcações da mudança de comportamento. As marcações para uma das tomadas da base UMN, obtidas ao final da enquete, são ilustradas na Figura 16. Os intervalos foram definidos ao avaliar a divergência nas marcações dos participantes.

Nessa figura, as áreas na cor verde – linhas na diagonal – correspondem ao comportamento normal da cena. O intervalo em amarelo – região em xadrez – descreve a área onde o comportamento anormal teve início. A região na cor vermelha – área pontilhada – é o decorrer do comportamento anormal. Por fim, a área em azul – linhas horizontais – é o intervalo que corresponde ao fim do comportamento anormal. Assim, os intervalos em amarelo e em azul são trechos onde ao menos um participante da enquete disse ser

Figura 16 – Marcações de uma das tomadas da base UMN com intervalos para comportamento normal e início, decorrer e fim do comportamento anormal, nas cores verde, amarelo, vermelho e azul, respectivamente.



Fonte: O autor (2016).

de comportamento anormal. Na região em vermelho, todos os participantes disseram ser referente a comportamento anômalo.

Essa forma de descrever os eventos nos vídeos foi adotada, pois como mostra Chaquet, Carmona e Fernández-Caballero (2013), rotular bases de vídeos por intervalos é uma prática eficaz para descrever os comportamentos na cena. Nas onze tomadas da base UMN, as pessoas simulam um comportamento semelhante ao visto na Figura 16. Os vídeos da base estão a uma taxa de 30 quadros por segundo, resolução de 320×240 e não apresentam perda visual de quadros no momento da aquisição – *drop frame*.

5.1.2 Métodos Utilizados

Para identificar os algoritmos baseados na aparência capazes de indicar mudança de comportamento em vídeos, foram realizados experimentos com seis métodos: MSM, PCA, 2D-PCA, IPCA, SSIM e IMED.

O SSIM é comumente aplicado para avaliar os níveis de perda de informação ao comprimir uma imagem e o IMED é utilizado para comparar caracteres e símbolos em um conjunto de figuras nas cores branco e preto. Os dois métodos foram incluídos nos testes devido à forma como eles avaliam a semelhança entre as imagens. O SSIM é baseado na estrutura da cena, como linhas, contornos ou pontos desenhados da imagem. Ele faz a comparação da estrutura das duas cenas e retorna um valor de similaridade entre as imagens.

O IMED é um método que identifica a distância entre as imagens e dá um valor para as diferenças entre cada pixel. Assim, é possível obter a distância entre as imagens avaliadas.

O SSIM e o IMED buscam um valor para representar a distância entre as imagens, porém o IMED não considera os dados de entrada como uma imagem, enquanto o SSIM foi desenvolvido especificamente para comparar imagens. Com isso, pequenas variações de iluminação fazem com que o IMED retorne grandes valores de distância, mesmo em cenas semelhantes. O método SSIM, não apresenta esta desvantagem, pois avalia a cena com base na estrutura apresentada na imagem.

Os métodos do subespaço PCA, 2D-PCA e IPCA fazem a classificação das imagens com base na distância entre as representações construídas para as cenas ao projetá-las em um subespaço. Nos experimentos realizados, foi utilizada a distância euclidiana para todos os métodos do subespaço. Em problemas de classificação de imagens baseados em métodos do subespaço, a distância euclidiana é uma das medidas mais utilizadas, pois apresenta bons resultados, apesar de sua simplicidade. O MSM, avalia o ângulo canônico entre as representações construídas por ele para relacionar os elementos projetados. Seja para a medida de distância euclidiana como para a medição baseada em ângulo canônico, valores menores representam maior confiança entre as cenas avaliadas.

5.1.3 Métricas de Avaliação dos Resultados

Os resultados obtidos na primeira série de experimentos são baseados nos valores de similaridade retornados pelos métodos MSM, PCA, 2D-PCA, IPCA, SSIM e IMED. As maiores diferenças indicam mudança na estrutura da imagem ou o aumento na distância entre os subespaços. Para determinar quais valores de similaridade representam mudança de comportamento é utilizada uma heurística de limiar auto-adaptado. O módulo marca os quadros do vídeo onde os níveis de similaridade apontam uma mudança de comportamento. Dessa forma, os quadros que não são marcados fazem referência ao comportamento normal e os que recebem marcação são rotulados como quadros com comportamento suspeito.

Como apresenta a Seção 5.1.1, as marcações de *ground truth* descrevem intervalos do vídeo. Dessa forma, a qualidade das marcações de cada método é dada ao verificar se os quadros marcados como suspeitos estão dentro do intervalo descrito como comportamento anormal ou não, o que contabiliza verdadeiro positivo (VP) ou falso positivo (FP), respectivamente. Caso um intervalo do *ground truth* marcado como comportamento anormal não receba marcação de comportamento suspeito, é contabilizado um falso negativo (FN). Assim, é possível estipular medidas de precisão e revocação dos métodos pelas fórmulas:

$$p = \frac{VP}{VP + FP} \qquad \text{e} \qquad r = \frac{VP}{VP + FN} \tag{5.1}$$

respectivamente. A seção a seguir exibe os resultados obtidos na primeira série de experimentos e discute sobre cada um dos seis métodos avaliados.

5.1.4 Resultados

Para os experimentos, foram definidos valores do parâmetro ω de 1 a 8. O parâmetro β do módulo de limiar auto-adaptado foi mantido em 9. A Tabela 5 mostra os níveis de precisão e revocação para os seis métodos, ordenados pelo valor da média dos resultados obtidos. Em destaque, os melhores níveis de precisão obtidos para cada valor do parâmetro ω.

Tabela 5 – Níveis de precisão e revocação para os seis métodos baseados na aparência avaliados na primeira série de experimentos.

Método	Parâmetro ω								
	1	2	3	4	5	6	7	8	
	Precisão								Média
2D–PCA	0.880	0.880	0.882	**0.860**	0.873	0.865	0.888	0.896	**0.878**
PCA	0.874	0.874	**0.883**	0.853	**0.875**	**0.881**	0.881	0.890	0.876
SSIM	**0.883**	**0.883**	0.841	0.833	0.813	0.862	**0.896**	**0.899**	0.864
IMED	0.836	0.836	0.776	0.788	0.864	0.880	**0.896**	0.882	0.845
IPCA	0.865	0.865	0.817	0.812	0.781	0.776	0.734	0.726	0.797
MSM	0.768	0.768	0.737	0.701	0.648	0.737	0.750	0.810	0.740
	Revocação								Média
2D-PCA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
PCA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SSIM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
IMED	0.989	0.989	0.975	0.943	0.924	0.964	0.962	0.971	0.965
IPCA	1.000	1.000	1.000	0.993	0.992	0.991	0.988	0.986	0.994
MSM	0.987	0.987	0.973	0.965	0.982	0.966	0.955	0.944	0.970

Fonte: O autor (2016).

Os piores níveis de precisão foram obtidos com o parâmetro $\omega = 4$ e $\omega = 5$. Este resultado está relacionado com características do vídeo, como taxa de quadros por segundo e níveis de ruído. A variação desses elementos pode alterar os resultados em relação aos valores do parâmetro ω. Os valores para $\omega = 1$ e $\omega = 2$ são iguais, como mostra a Tabela 5. Este comportamento ocorre devido à baixa alteração na estrutura da cena ao combinar dois quadros adjacentes em uma única imagem média, quando $\omega = 2$. Neste caso, o tempo necessário para avaliar o vídeo com $\omega = 2$ é menor em relação ao tempo para $\omega = 1$, pois apenas uma similaridade é calculada para quatro imagens, isto é, dois quadros para cada imagem média. Assim, o aumento do valor do parâmetro ω reduz o tempo necessário para avaliar o vídeo. O tempo de execução dos métodos é tratado em detalhes na Seção 5.2.4.

O método 2D-PCA, em destaque na Tabela 5, apresentou a melhor média de precisão e revocação máxima, enquanto que as abordagens IPCA e MSM apresentam as piores taxas de precisão. Os valores de similaridade gerados pelos dois métodos são semelhantes aos valores calculados pelo 2D-PCA. Assim, o módulo de limiar auto-adaptado pode ser remodelado para que as abordagens sejam tratadas de forma condizente com seus valores de similaridade apresentados e, dessa forma, melhorar os resultados em abordagens como IPCA e MSM. Essa discussão é tratada em detalhes na Seção 5.3.

O IMED apresentou a menor revocação e um dos piores níveis de precisão. Sua abordagem clássica pode não representar de forma adequada as mudanças na cena, pois o método não considera a relação espacial dos elementos na imagem. Dessa forma, o IMED não será aplicado na próxima série dos experimentos. Os métodos SSIM, PCA e 2D-PCA apresentaram as maiores taxas de precisão e tiveram revocação máxima, pois não deixaram de alertar comportamento anormal em nenhuma cena. Nesse tipo de sistema de segurança, um falso positivo não é tão grave quanto um falso negativo. Assim, foram selecionados para a próxima série de experimentos os métodos: SSIM, PCA e 2D-PCA. A próxima seção trata dos experimentos com a arquitetura proposta e avalia os tempos de execução de cada fase do método.

5.2 Experimento Com a Arquitetura Proposta

A segunda série de experimentos avalia toda a arquitetura em três bases de dados: UMN, descrita na Seção 5.1.1; PETS *Crowd Sensing Dataset Challenge* (FERRYMAN; SHAHROKNI, 2009),[2] (PETS2009) e *YouTube Abnormal Behavior* (YAB), descritas em detalhes da Seção 5.2.1. A Seção 5.2.2 fala sobre os métodos que compõem a fase baseada em características. Na Seção 5.2.3 é descrita a métrica de avaliação dos resultados. Em seguida, a Seção 5.2.4 apresenta os resultados obtidos nessa série de experimentos, avalia o custo de computação de cada uma das fases da proposta, discorre sobre a saída gerada na fase baseada em características e compara o método proposto com outras duas abordagens encontradas na literatura.

5.2.1 Bases de Dados Investigadas

A base PETS2009 é um conjunto de vídeos que trata vários tipos de problemas que envolvem comportamento de pessoas em locais públicos. Os vídeos são acompanhados de marcações para os quadros onde ocorrem as mudanças de comportamento, além de outras informações como o número de pessoas na cena e o sentido do deslocamento. No entanto, para os testes realizados, as informações do quadro onde ocorre a mudança de comportamento foram modificadas para o formato que descreve intervalos de comportamento normal e início, decorrer e fim do comportamento anormal. As marcações foram realizadas no mesmo procedimento descrito na Seção 5.1.1.

A base é dividida em quatro sub-conjuntos, chamados de S0, S1, S2 e S3, Este último, trata de comportamento em cenas com multidão. Foram selecionados 12 vídeos para os experimentos, com 4 que simulam dispersão e 8 com simulação semelhante à apresentada

[2] PETS2009 S3. Disponível em: <ftp://ftp.pets.rdg.ac.uk/pub/PETS2009/Crowd_PETS09_dataset/a_data/Crowd_PETS09/S3_HL.tar.bz2>.

na base UMN. A Figura 17 mostra os quatro enquadramentos presentes no sub-conjunto S3 da base PETS2009.

Figura 17 – Os quatro enquadramentos presentes na base PETS2009.

Fonte: Adaptado de Ferryman e Shahrokni (2009).

Os vídeos da base PETS2009 presentam perda de quadros no momento da aquisição – *drop frame* – e a taxa de quadros por segundo está em 15 quadros com o total de 62 segundos de duração para os 12 vídeos. A resolução dos vídeos é de 720×480.

A base *YouTube Abnormal Behavior* foi criada para avaliar o método proposto neste trabalho durante os experimentos. A base é formada por 52 vídeos de sistemas de monitoramento reais baixados do site *Youtube*[3]. Os vídeos apresentam vários tipos de comportamento anormal, como: brigas, tiroteios e assaltos. Para os experimentos com a base YAB foram selecionados 20 vídeos que apresentam comportamento semelhante ao que é simulado nas bases UMN e PETS2009. A Figura 18 mostra algumas cenas da base YAB.

Figura 18 – Imagem de três vídeos que compõem a base YAB.

Fonte: O autor (2016).

As marcações foram realizadas por meio da enquete descrita na Seção 5.1.1 e dividem o comportamento nos vídeos por intervalos. Os 20 vídeos apresentam taxas de quadros por segundo entre 15 e 30 quadros, com tempo total de 11 minutos e resoluções variadas, que vão de 320×240 a 854×480. A maioria dos vídeos está em baixa qualidade, o que resulta em alta ocorrência de artefatos gerados pelo codificador de vídeo no processo de compressão.

5.2.2 Métodos Utilizados

Para esta série de experimentos a arquitetura proposta foi aplicada com todos os módulos descritos no Capítulo 4. O funcionamento dos módulos da fase baseada na aparência é o

[3] Site YouTube: <http://www.youtube.com/>.

mesmo apresentado na Seção 5.1. No entanto, foram utilizados somente três métodos para gerar os valores de similaridade, como descreve a Seção 5.1.4, são eles: SSIM, PCA e 2D-PCA.

Sempre que um trecho do vídeo é marcado como suspeito, a fase baseada em características calcula o fluxo óptico para o par de imagens médias que gerou o sinal. O resultado é um conjunto de descritores do fluxo óptico estimado para aquele momento do vídeo, na forma de vetores sob o plano bidirecional da imagem. Esse conjunto de características é fornecido com entrada para o módulo responsável por avaliar o fluxo óptico com o operador de divergência.

O operador de divergência gera um mapa com valores de divergência/convergência. Esse mapa é dividido em células, como descreve no Capítulo 4. As células com convergência são marcadas como aglomeração, e as com divergência recebem o rótulo de dispersão. As demais células são rotuladas como normais. Caso a quantidade de células normais seja inferior a um limiar, o comportamento suspeito é marcado como anormal. Caso contrário, a fase baseada em características é desativada.

Por fim, se o comportamento é confirmado como anormal, as regiões das células são utilizadas para marcar a imagem do vídeo e exibir para o operador humano as áreas com comportamento anormal, conforme seu rótulo de aglomeração ou dispersão.

5.2.3 Métricas de Avaliação dos Resultados

Nesta série de experimentos, é avaliada a qualidade das marcações geradas pelo método proposto e seu tempo de execução. As fases baseada na aparência e baseada em características são avaliadas separadamente quanto ao tempo de execução. Dessa forma, é possível calcular a redução do custo de computação ao dividir o sistema inteligente em duas fases e manter somente parte da arquitetura em execução durante cenas de comportamento normal. Os tempos de cada fase são comparados com os tempos de execução dos métodos apresentados por Gu, Cui e Zhu (2014) e Liu, Li e Jia (2014).

A qualidade das marcações é dada pela quantidade de alarmes disparados dentro do intervalo descrito como comportamento anormal, segundo o *ground truth*. As medidas de precisão e revocação utilizadas na Seção 5.1 não se aplicam a essa série dos experimentos. Isso ocorre dada a forma como os métodos geram suas marcações: baseada em alarmes disparados no decorrer do vídeo, sem realizar marcações quadro-a-quadro.

Os alarmes são contabilizados para cada vídeo da base. A quantidade de acertos é dada para o número de vídeos que o método teve um alarme disparado em áreas que descrevem comportamento anormal. No entanto, para identificar atraso nas marcações, em relação ao início do comportamento, são considerados dois tipos de acerto: (1) o alarme disparado no intervalo de início do comportamento anormal, exibido na Figura 16 na cor amarela, e (2) o alarme disparado para o restante do intervalo da base marcado como

comportamento anormal, que na Figura 16 são os intervalos em vermelho e azul.

Esses alarmes, chamados de marcações com atraso, não consistem em erro, pois são referentes ao comportamento anormal corrente. No entanto, é dito que o método apresentou atraso em sua marcação pois seu alarme está após a marca mais tardia gerada pelos entrevistados ao tentarem marcar o que, visualmente, é o início do comportamento anormal. As marcações nos intervalos descritos como comportamento normal contabilizam falso positivo. São utilizadas nos experimentos as bases UMN, PETS2009 e os 20 vídeos da base YAB selecionados, como descrito na Seção 5.2.1.

5.2.4 Resultados Obtidos

Conforme mencionado anteriormente, o método proposto nesta pesquisa foi comparado às propostas apresentadas por Liu, Li e Jia (2014) e Gu, Cui e Zhu (2014). Os três métodos tiveram seus parâmetros ajustados de forma que suas respectivas sensibilidades na identificação dos comportamentos anormais apresentassem os melhores resultados para cada uma das bases.

A abordagem descrita por Liu, Li e Jia (2014), assim como a proposta apresentada neste trabalho, não passa por uma etapa de aprendizagem. O método utiliza uma heurística de limiar auto-adaptado própria que avalia as características extraídas da cena ao calcular o fluxo óptico esparso. Na proposta apresentada por Gu, Cui e Zhu (2014), um classificador GMM é utilizado para identificar quadros com taxas de dispersão e intensidade de movimentos. O método passa por uma etapa inicial de treino para definir os parâmetros do GMM e utiliza a abordagem de vizinhos próximos para classificar novos quadros como comportamento normal ou anormal.

Os métodos foram avaliados quanto à qualidade das marcações de acordo com o momento em que os alarmes são disparados para a linha de tempo dos vídeos. As marcações podem ocorrer no início do comportamento anormal, com atraso ou falsos positivos. A Tabela 6 apresenta a quantidade de vídeos da base UMN com alarmes disparados no intervalo onde ocorre um comportamento anormal e a quantidade de falsos positivos para toda a base pelos três métodos. Os valores entre parênteses são o percentual aproximado que o número de acertos representa em relação à quantidade de vídeos na base de dados. A base UMN é composta por 11 vídeos.

Todas as abordagens dispararam alarmes no momento em que algum comportamento anormal estava presente da cena. No entanto, os métodos propostos por Liu, Li e Jia (2014) e Gu, Cui e Zhu (2014) geraram grande parte de suas marcações com atraso em relação ao início do comportamento anormal. Nesse experimento, os resultados do método proposto foram os mesmos ao utilizar SSIM e PCA nos módulos de análise da similaridade. Os resultados com o uso do 2D-PCA foram piores, em relação aos métodos SSIM e PCA.

Tabela 6 – Alarmes disparados pelos três métodos para a base UMN.

Método	Tipo de marcação			
	Início	Com atraso	Sem alarmes	Falso positivo
(LIU; LI; JIA, 2014)	5 (46%)	6 (54%)	0	42
(GU; CUI; ZHU, 2014)	6 (54%)	5 (46%)	0	13
Método proposto*	9 (82%)	2 (18%)	0	17

* Com SSIM e PCA na fase baseada na aparência.

Fonte: O autor (2016).

A Tabela 7 exibe os resultados obtidos para os experimentos na base PETS2009. Vídeos que não obtiveram alarmes disparados por nenhuma abordagem são contabilizados na terceira coluna de resultados obtidos. Durante o ajuste de parâmetros para adequar a sensibilidade dos métodos para os vídeos sem marcação, foi observada a redução na qualidade das marcações nos demais vídeos. Portanto, foi mantido o conjunto de parâmetros com o melhor resultado geral de cada método. Entre parênteses, o percentual aproximado que o número de acertos representa em relação à quantidade de vídeos na base de dados. Como descreve a Seção 5.2.1, foram utilizados nos experimento 12 vídeos da base PETS2009.

Tabela 7 – Alarmes disparados pelos três métodos avaliados na base PETS2009.

Método	Tipo de marcação			
	Início	Com atraso	Sem alarmes	Falso positivo
(LIU; LI; JIA, 2014)	2 (17%)	6 (50%)	4 (33%)	3
(GU; CUI; ZHU, 2014)	8 (66%)	3 (26%)	1 (8%)	15
Método proposto*	3 (26%)	7 (58%)	2 (16%)	5

* Com SSIM na fase baseada na aparência.

Fonte: O autor (2016).

Os resultados das marcações mostram que a abordagem apresentada por Gu, Cui e Zhu (2014) tem bons resultados na base PETS2009, que apresenta vídeos com altos níveis de *drop frame* e baixa taxa de quadros por segundo. No entanto, a estratégia apresentou a maior quantidade de falsos positivos. Os resultados referentes ao método proposto são para o uso do SSIM na fase baseada na aparência. Ao utilizar o PCA e o 2D-PCA, as marcações apresentaram qualidade inferior, em relação ao uso do SSIM.

Os resultados obtidos nos experimentos com a base YAB são exibidos na Tabela 8. Como a base apresenta vídeos com maior duração, em relação às bases UMN e PETS2009, o tamanho do conjunto de treino do método proposto por Gu, Cui e Zhu (2014), que não foi definido como um parâmetro, foi ajustado no intuito de melhorar a qualidade das marcações desta abordagem. Os valores descritos entre parênteses são o percentual aproximado do número de acertos apresentado, em relação aos 20 vídeos da base YAB utilizados nos experimentos.

Os resultados apresentam a abordagem descrita por Liu, Li e Jia (2014) como a única que não disparou alarmes para todos os vídeos da base YAB. Nesse experimento, assim como no anterior, o método proposto apresentou os melhores resultados com o uso do SSIM para calcular os valores de similaridade. A abordagem apresentada por Gu, Cui e Zhu (2014)

Tabela 8 – Alarmes disparados pelos três métodos para a base YAB.

Método	Tipo de marcação			
	Início	Com atraso	Sem alarmes	Falso positivo
(LIU; LI; JIA, 2014)	12 (60%)	4 (20%)	4 (20%)	55
(GU; CUI; ZHU, 2014)	17 (85%)	3 (15%)	0	50
Método proposto*	18 (90%)	2 (10%)	0	73

* Com SSIM na fase baseada na aparência.

Fonte: O autor (2016).

apresentou o menor índice de falsos positivos, mas teve um número levemente menor de marcações para o início do comportamento anormal, em relação ao método proposto.

Quanto ao tempo de execução dos três métodos, foram avaliados somente os intervalos onde são executados os procedimentos que compõem a arquitetura de cada um. O tempo necessário para acessar os vídeos em disco ou para exibir marcações na tela, se for o caso, não é contabilizado. Os experimentos foram executados em um computador com 8GB de RAM, processador de 2.3GHz e sistema operacional Windows 64 Bits. Os três métodos utilizam módulos desenvolvidos em C++ com bibliotecas do OpenCV 3.0 e recursos presentes no MATLAB.

A contagem de tempo das abordagens é feita de duas formas: (1) tempo total gasto para analisar cada uma das bases investigadas e (2) tempo da análise feita pelas propostas para cada trecho do vídeo. Os resultados exibidos são a média de tempo para 12 execuções. O tempo de execução do método proposto apresenta três medidas, uma para cada tipo de abordagem utilizada na fase baseada na aparência ao calcular os valores de similaridade, são elas: SSIM, PCA e 2D-PCA. A Tabela 9 apresenta o tempo total de execução, em segundos, dos métodos avaliados nas três bases investigadas.

Tabela 9 – Tempo médio de execução dos métodos ao avaliar toda a base.

Método	Tempo total de execução (s)		
	UMN	PETS2009	YAB
(LIU; LI; JIA, 2014)	39.50	29.18	457.31
(GU; CUI; ZHU, 2014)	148.64	98.59	1786.17
Método proposto (SSIM)	33.25	17.09	319.57
Método proposto (PCA)	22.81	8.49	168.73
Método proposto (2D-PCA)	73.16	53.90	1066.74

Fonte: O autor (2016).

A abordagem apresentada por Gu, Cui e Zhu (2014) tem os maiores tempos, pois calcula o Fluxo Óptico denso para todos os trechos do vídeo, o que implica em alto custo de computação. Além disso, o método cria um novo modelo GMM para cada vídeo ao alcançar a fase de treino, que consiste no momento onde um dado número de quadros é lido durante sua execução. Para os tempos exibidos na Tabela 9, o tempo médio de treino do método descrito por Gu, Cui e Zhu (2014) é 9.1, 1.9 e 5.9 milissegundos para as bases UMN,

PETS2009 e YAB respectivamente. O método proposto aplica o fluxo óptico denso apenas nos trechos apresentados como suspeitos pela fase baseada na aparência, assim, o tempo gasto para avaliar o vídeo é menor. Ao utilizar o SSIM e o PCA na fase baseada na aparência, o método proposto apresentou tempos menores que as demais propostas avaliadas.

A Tabela 10 apresenta os tempos, em milissegundos, da análise feita pelos métodos para cada trecho do vídeo. A proposta deste trabalho apresenta duas medidas de tempo: da fase baseada na aparência e da baseada em características. Dessa forma, é possível identificar o tempo médio do método na fase baseada na aparência ao avaliar cenas com comportamento normal ou que apresentem apenas a imagem de fundo.

Tabela 10 – Tempo médio de execução para cada trecho dos vídeos. O método proposto apresenta tempos para as fases baseada na aparência (Ap.) e baseada em características (Car).

Método	Tempo de execução para cada trecho do vídeo (ms)					
	UMN		PETS2009		YAB	
(LIU; LI; JIA, 2014)	5.8		5.2		29.9	
(GU; CUI; ZHU, 2014)	19.2		99.5		93.1	
	Ap.	Car.	Ap.	Car.	Ap.	Car.
Método proposto (SSIM)	7.1	6.3	31	15.8	28.8	13.7
Método proposto (PCA)	3.2	6.2	10.4	14.9	10.9	18.6
Método proposto (2D-PCA)	15.8	6.7	104.4	19.5	105.4	19.3

Fonte: O autor (2016).

O método apresentado por Liu, Li e Jia (2014) apresenta os menores tempos de execução na base PETS2009, dado o uso da abordagem esparsa do Fluxo Óptico. Entretanto, a estratégia apresentada por Gu, Cui e Zhu (2014), baseada no fluxo óptico denso, se mostrou muito sensível ao aumento da resolução do vídeo, visto que seus piores tempos foram para as bases YAB e PETS2009. O método proposto apresentou tempos inferiores à 33.3 milissegundos, para todas as contagens de tempo ao utilizar os métodos SSIM e PCA na fase baseada na aparência. Esse tempo é o menor intervalo entre dois quadros de vídeo para as bases utilizadas nos experimentos e é referente aos vídeos com taxas de 30 quadros por segundo. Assim, executar análises num intervalo de tempo inferior à 33.3 milissegundos possibilita que o método proposto seja capaz de avaliar os vídeos em tempo real, isto é, sem inferir atraso na execução do sistema ou na exibição das imagens.

Durante os experimentos, o comportamento das duas fases que compõem o método proposto foi analisado separadamente. A fase baseada em características, como descrito na Seção 4, deve identificar alarmes falsos disparados pela fase baseada na aparência. Assim, ao avaliar os alarmes da fase baseada na aparência, foi constatado que cerca de 30% dos sinais disparados nas três bases de dados foram falsos alarmes identificados pela fase baseada em características. Com isso, o método para reconhecer comportamento em vídeos composto somente pela fase baseada na aparência dispara, em alguns vídeos, até

duas vezes mais alarmes falsos do que o método composto por toda a arquitetura proposta.

A fase baseada em características, quando executada sem a fase baseada na aparência e sem esperar por um sinal de ativação, apresentou um número maior de falsos positivos, após novo ajuste de parâmetros, mas mostrou comportamento semelhante ao da arquitetura completa em relação à qualidade das marcações com acerto. Quanto ao tempo da avaliação, foi observado que, executar a fase baseada em características durante todo o vídeo apresenta sempre um tempo maior em relação ao tempo de execução da arquitetura completa. O aumento no tempo de execução é de aproximadamente 38% nas bases de dados PETS2009 e YAB. Para a base UMN, com menor dimensão nas imagens, o aumento do tempo de execução chega a 15%. Dessa forma, o uso da duas fases juntas no método proposto reduz a quantidade de falsos positivos e o tempo de execução ao avaliar os vídeos, quando comparado às fases executadas separadamente.

A qualidade das marcações geradas pelo método proposto para auxiliar o operador humano foram avaliadas visualmente, pois as bases de dados utilizadas nos experimentos não possuem rótulos que descrevem o momento e as regiões na imagem onde um comportamento de dispersão ou aglomeração ocorre. A Figura 19 mostra três cenas de vídeos da base YAB utilizados nos experimentos. As imagens estão marcadas com blocos gerados pelo método proposto e descrevem regiões de aglomeração, em vermelho, e de dispersão, na cor azul.

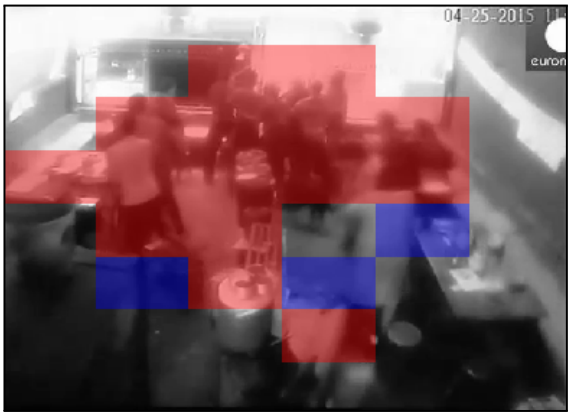
A Figura 19 (a) exibe uma cena onde as pessoas saem dos prédios e das calçadas em direção ao centro da rua, após o início de um terremoto. As regiões em vermelho mostram a área na imagem onde as pessoas se reúnem. Na Figura 19 (b), as pessoas na cena correm em direção à saída do estabelecimento, região em vermelho no topo do imagem. Os blocos em azul mostram de onde as pessoas partem, assim que o comportamento anormal tem início. Por fim, a Figura 19 (c) mostra um cenário onde as pessoas fogem para dentro de uma loja, à esquerda do vídeo. A área em vermelho é em frente à entrada do estabelecimento onde as pessoas se agrupam durante o comportamento anômalo.

Nas três cenas, é possível ver o deslocamento das pessoas em direção ao centro das áreas em vermelho, que indicam comportamento anormal de aglomeração. Os blocos em azul, que marcam regiões com dispersão, mostram de onde as pessoas partem, após o início do comportamento anômalo.

As marcações na base UMN são ilustradas na Figura 20 em duas tomadas. Como descrito na Seção 5.1.1, a base de dados apresenta o mesmo tipo de comportamento nas onze tomadas, onde as pessoas se deslocam no centro da imagem e, em seguida, correm para fora da cena, ao simular um cenário de dispersão. A Figura 20 (a) mostra o início do comportamento anormal reconhecido pelo método proposto em uma das tomadas da base UMN. Os blocos em azul indicam a dispersão das pessoas do centro da cena para a parte inferior da imagem. Os blocos em vermelho marcam as áreas da cena para onde as pessoas se movimentam durante a simulação de evacuação.

Figura 19 – Três vídeos da base YAB com marcações de dispersão e aglomeração geradas pelo método proposto, nas cores vermelho e azul, respectivamente.

(a) Vídeo 07 da base YAB: pessoas se aglomeram nas ruas (marcações em vermelho) após o início de um terremoto.



(b) Vídeo 15 da base YAB: pessoas se aglomeram na saída de um estabelecimento (marcações em vermelho) após o início de um terremoto.

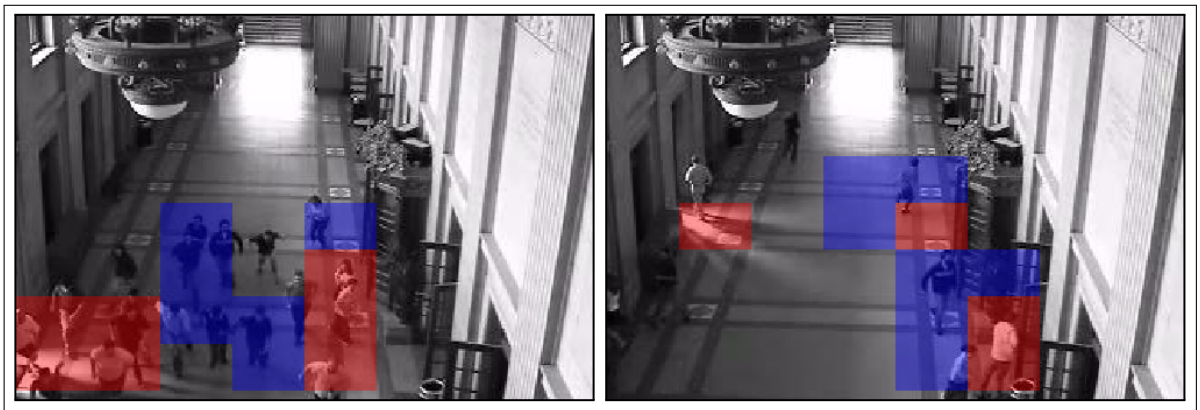
(c) Vídeo 17 da base YAB: pessoas se aglomeram na entrada de um estabelecimento (marcações em vermelho) após o início de um tiroteio.

Fonte: O autor (2016).

Na Figura 20 (b), o mesmo tipo de comportamento é marcado em outra tomada da base UMN, mas com uma imagem do final do comportamento anormal. Os blocos em azul indicam a dispersão das últimas pessoas presentes na cena. As áreas em vermelho são as regiões de escape, marcadas como pontos de aglomeração. A imagem mostra que o método proposto mantém sua análise do comportamento durante todo o vídeo. Caso a anormalidade permaneça, as marcações são atualizadas para o operador humano. Assim que o comportamento volta ao normal as marcações são removidas e o método proposto continua a análise do vídeo com a fase baseada na aparência em busca de novo comportamento anormal.

Os vídeos selecionados para os experimentos com a base PETS2009 apresentam dois tipos de comportamento anormal, como descreve a Seção 5.2.1. Em um desses, as pessoas na cena correm em uma única direção. No outro comportamento, todos estão parados no centro da imagem e a anormalidade tem início quando as pessoas se dispersam. Nos

Figura 20 – Duas tomadas da base UMN com marcações de dispersão e aglomeração geradas pelo método proposto, nas cores vermelho e azul, respectivamente.

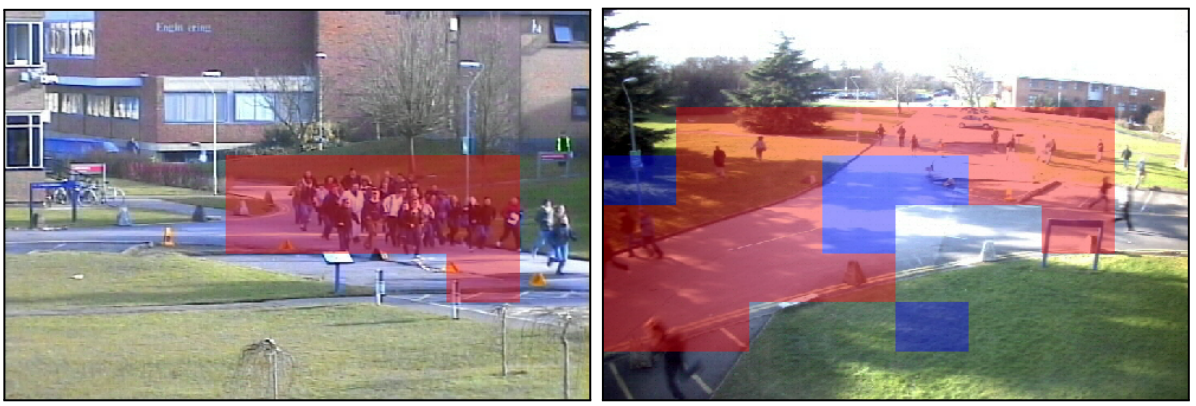


(a) Início do comportamento anormal. (b) Fim do comportamento anormal.

Fonte: O autor (2016).

experimentos foi observado que, nas cenas onde todas as pessoas se movem na mesma direção sem apresentar aglomeração ou dispersão durante o comportamento anormal, o método gera marcações apenas em vermelho, de aglomeração, como mostra a Figura 21 (a). Isso ocorre pois, para esse tipo de deslocamento, o operador ∇, descrito na Seção 4.1.2, retorna valores de convergência – aglomeração – maiores que os de divergência, referentes à dispersão. Como resultado, cenas com esse tipo de movimento não apresentam marcações referentes aos valores de dispersão – em azul – que estão localizados no entorno da área de convergência.

Figura 21 – Cenas da base PETS2009 com marcações de dispersão em azul e aglomeração, na cor vermelha, geradas pelo método proposto.



(a) Simulação onde todos na imagem correm na mesma direção.

(b) Simulação onde todos na imagem de dispersam a partir do centro da cena.

Fonte: O autor (2016).

Na Figura 21 (b), a área marcada em azul, ao centro da imagem, sinaliza a regão de onde as pessoas se dispersaram. As marcações em azul próximas à borda da imagem indicam que alguém saiu da cena. As áreas em vermelho, de aglomeração, marcam as regiões na cena onde as pessoas correm durante o comportamento anormal.

Foi observado que, visualmente, as marcações geradas pelo método proposto para auxiliar o operador humano descrevem de forma adequada o comportamento presente nos vídeos das três bases de dados investigadas. O tempo de execução do método não é comprometido ao inserir as marcações visuais no vídeo, pois são necessários no máximo 2,5 milissegundos para realizar essa tarefa, mesmo na maior resolução de vídeo presente nos experimentos.

As pesquisas apresentadas por Gu, Cui e Zhu (2014) e Liu, Li e Jia (2014) não geram marcações visuais. Elas marcam o vídeo apenas em escopo global, ou seja, disparam alarmes que indicam o tempo onde o comportamento anormal foi identificado. O método proposto, no entanto, atua em escopo global e local, pois indica o tempo do vídeo onde o comportamento anômalo foi identificado e exibe marcações visuais que descrevem o local na imagem onde ele ocorre. As marcações, por sua vez, informam quais regiões da imagem são de comportamento anormal do tipo aglomeração e quais são do tipo dispersão.

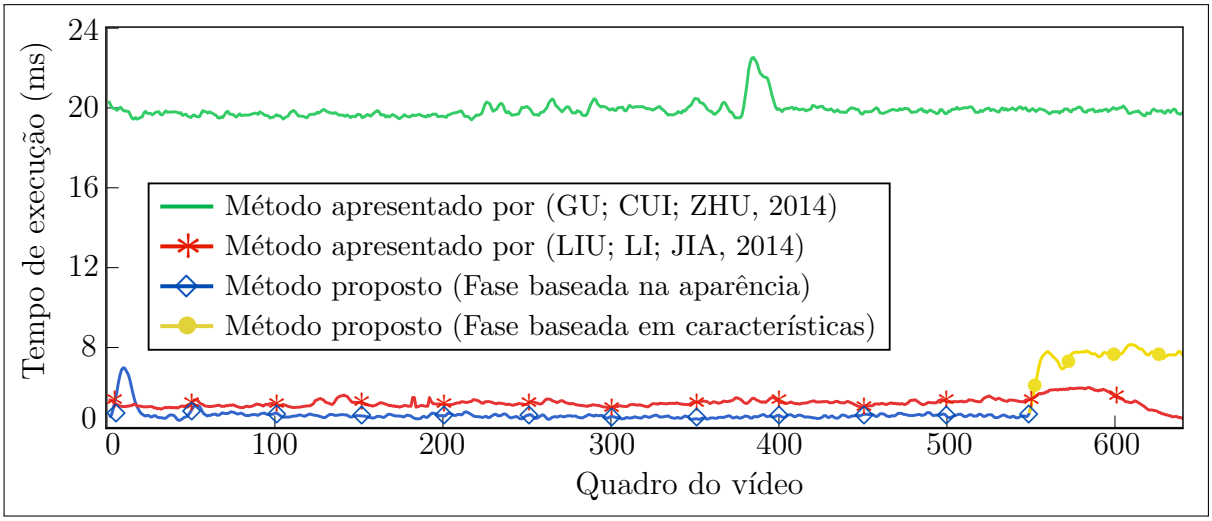
5.3 Considerações Finais

Os experimentos realizados mostram que a proposta apresentada é capaz de reconhecer mudança de comportamento na cena. O método proposto foi comparado a outros dois. Ele disparou sinais que indicam o reconhecimento de um comportamento anômalo com qualidade equivalente ou superior às demais abordagens. Como resultado, a abordagem marca regiões na imagem para auxiliar o operador humano a identificar o local e o tipo da anormalidade presente na cena. Uma arquitetura modular é apresentada e propõe o uso de partes específicas do sistema para cada tipo de comportamento. Ao avaliar o tempo de execução do método proposto, foi identificado que a arquitetura possibilita a redução do tempo global de execução do sistema. Ao comparar o desempenho dos métodos SSIM, PCA e 2D-PCA na segunda série de experimentos, foi constatado que a melhor configuração para o método proposto é com o método SSIM na fase baseada na aparência. A qualidade da detecção de comportamento anormal pelo método proposto ao utilizar esse é superior, em relação ao uso do PCA e do 2D-PCA. O tempo de execução do SSIM, no entanto, é maior que o tempo de execução do PCA, mas é inferior ao tempo dos demais métodos investigados.

Durante a revisão de literatura, constatou-se que os métodos SSIM, IMED, MSM, PCA, IPCA e 2D-PCA não são comumente utilizados para identificar mudanças de comportamento em vídeos. No entanto, os resultados obtidos nos experimentos demonstram que a análise estrutural da cena ao utilizar esses métodos apresenta bons resultados ao identificar mudanças de comportamento em vídeos simulados e em cenas reais. Uma nova heurística de limiar auto-adaptado foi proposta. Seu uso nos experimentos para rotular os valores de similaridade da fase baseada na aparência apresentou bons resultados, apesar de sua simplicidade.

A arquitetura modular proposta neste trabalho possibilita controlar partes do sistema para execução em momentos específicos. Quando módulos são desativados para determinados trechos do vídeo, o tempo de computação do sistema é reduzido. Este comportamento pode ser observado na Figura 22, que exibe o tempo médio de execução das abordagens avaliadas para cada trecho de um dos vídeos da base UMN. Os tempos do método proposto são exibidos em dois intervalos distintos, nas cores azul e amarelo, para a fase baseada na aparência e baseada em características, respectivamente.

Figura 22 – Tempo médio de execução para cada trecho de um dos vídeos avaliados.



Fonte: O autor (2016).

Como mostra a Figura 22, durante a fase baseada na aparência o tempo de execução do método proposto é inferior ao tempo da abordagem apresentada por Liu, Li e Jia (2014). No entanto, a fase baseada em características, de maior custo, é ativada somente no momento onde um comportamento suspeito é identificado pelo módulo de limiar auto-adaptado que compõe a fase baseada na aparência. Os experimentos mostram que, apesar do baixo custo computacional do método apresentado por Liu, Li e Jia (2014), a qualidade das marcações geradas por ele nas bases PETS2009 e YAB são inferiores aos resultados obtidos pelo método proposto e também pela abordagem apresentada por Gu, Cui e Zhu (2014).

O número de falsos positivos disparados pelo método proposto aumenta de acordo com o nível de ruído presente nos vídeos. A base YAB, onde a proposta apresenta a maior quantidade de falsos positivos, tem baixa qualidade na maioria dos vídeos, o que resulta no surgimento de artefatos gerados pela compressão do vídeo. Esses artefatos modificam a estrutura da cena e dificultam o reconhecimento do comportamento. A Figura 23 (a) mostra o quadro de um dos vídeos da base YAB com uma região coberta por artefatos. A Figura 23 (b) exibe o mesmo quadro com a região em destaque e a Figura 23 (c) exibe o quadro seguinte, sem os artefatos na mesma região.

O ruído apresentado na forma de artefatos é prejudicial para o comportamento do método proposto. Na fase baseada na aparência, como o módulo responsável por calcular

Figura 23 – (a) quadro do vídeo com região coberta por artefatos gerados no momento de compressão do vídeo, (b) região com artefatos em destaque para o mesmo quadro e (c) quadro seguinte sem os artefatos na mesma região.

(a) (b) (c)

Fonte: O autor (2016).

os valores de similaridade é sensível às mudanças na estrutura da imagem, o número de quadros marcados como suspeitos é maior. Na fase baseada em características, o Fluxo Óptico ameniza os problemas causados por esse tipo de artefato, pois esse tipo de ruído não se desloca no espaço da cena, o que possibilita identificar falsos positivos por parte da fase baseada na aparência. No entanto, o fluxo óptico gera descritores para os artefatos no momento em que eles desaparecem. Nesse instante, falsos positivos são contabilizados para o método. Uma solução para este problema pode ser alcançada ao aplicar uma extensão do Fluxo Óptico, como a proposta LDOF, apresentada na Seção 3.

O próximo capítulo apresenta, na Seção 6.1, as conclusões obtidas nesta pesquisa e em seguida trata dos tópicos que servirão de base para direcionar os trabalhos futuros, na Seção 6.2.

6 Conclusões e Trabalhos Futuros

Este capítulo trata, de forma sucinta, dos principais tópicos abordados no decorrer da pesquisa no sentido de apresentar as conclusões e os trabalhos futuros.

6.1 Conclusões

Este trabalho apresentou uma proposta de arquitetura modular e um método baseado nessa arquitetura capaz de reconhecer comportamento anormal em vídeos de multidão. A arquitetura, por sua vez, viabiliza o uso desse sistema inteligente em problemas reais, ao particionar o processo de reconhecimento em módulos independentes. O primeiro módulo, de menor custo computacional, é a fase baseada na aparência, responsável por identificar quando o vídeo apresenta uma cena de comportamento anormal. O segundo módulo é a fase baseada em características, que permanece inoperante até que um cenário de anomalia tenha início. Como o segundo módulo é o de maior custo computacional do método proposto, seu uso é interrompido sempre que um cenário normal está presente no vídeo.

Um dos pontos chave da pesquisa consistiu em combinar ferramentas capazes de atacar porções distintas do mesmo problema. Outras contribuições desta pesquisa são: (1) a abordagem baseada na análise da similaridade estrutural da cena para gerar índices de mudança do comportamento no vídeo, (2) uma heurística de limiar auto-adaptado para rotular os níveis de similaridade e (3) o uso de um operador de divergência aplicado aos descritores de fluxo óptico para identificar cenas de aglomeração e dispersão em vídeos de multidão.

Os experimentos mostraram que a arquitetura modular reduz o custo de computação ao analisar, de forma simplificada, o conteúdo do vídeo para determinar se há necessidade de ativar as demais partes do sistema, conforme sua função. O método proposto foi comparado com outros dois métodos: um baseado em uma heurística própria de limiar auto-adaptado e outro com um classificador treinado no início de cada experimento. Os resultados mostram que o tempo total de execução do método proposto é menor, em relação aos demais. Na função de reconhecer o comportamento na cena, a qualidade das marcações de comportamento anormal do método proposto é equivalente ou superior às marcações dos demais, inclusive com resultados que superam o baseline que utiliza aprendizagem para cada cena avaliada.

6.2 Trabalhos Futuros

A arquitetura e o método proposto neste trabalho são validados em suas respectivas tarefas. No entanto, problemas e limitações estão presentes nas ferramentas propostas e

servirão de guia para os trabalhos futuros. O primeiro tópico diz respeito à arquitetura. Sua descrição para um sistema de vigilância com múltiplas câmeras necessita de um protocolo que descreva seu funcionamento em ambientes paralelizáveis ou distribuídos. Ainda, são necessárias diretrizes que controlem a comunicação ente os módulos, para que o método não sobrecarregue o sistema ou o operador humano com inúmeras marcações e alarmes.

Para o módulo de análise de similaridade, novos métodos de análise da cena devem ser estudados, por exemplo, o *Fast* SSIM, extensão do SSIM que apresenta tempo de execução menor em relação a abordagem original. Devem ser avaliados métodos que atenuem problemas com variação de iluminação e ruídos que degradam a qualidade dos níveis de similaridade. A heurística de limiar auto-adaptado deve ser aprimorada, de forma que os rótulos dos níveis de similaridade sejam mais coesos e confiáveis. O uso de outras abordagens que tratam o ajuste automático de limiar deve ser investigado, como a proposta apresentada por Du et al. (2013), que trata dados ruidosos em uma rede de sensores e utiliza dois limiares auto-adaptados ao avaliar os dados recebidos como entrada.

Na fase baseada em características, extensões do Fluxo Óptico devem ser investigadas em suas abordagens densa e esparsa para reduzir o tempo de computação e sanar problemas gerados pela baixa qualidade dos vídeos, que resulta em artefatos. Devem ser investigados métodos capazes de derivar descritores para a cena e que possam substituir o fluxo óptico. O operador de divergência pode ser aprimorado ou substituído por um classificador. Assim, a análise da cena pode resultar em descrições mais detalhadas e com uma quantidade maior de tipos de comportamento identificado. Essas mudanças podem aumentar o custo computacional do método, fator que deve sempre ser levado em consideração ao trabalhar com aplicações que envolvam risco à integridade das pessoas e cuja eficácia esteja atrelada à uma resposta em tempo real por parte do sistema inteligente (VISHWAKARMA; AGRAWAL, 2012).

Referências

ALI, S.; SHAH, M. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: *proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Conference on CVPR*. Minnesota, USA: IEEE, 2007. p. 1 – 6.

ALLAIN, P.; COURTY, N.; CORPETTI, T. AGORASET: a dataset for crowd video analysis. In: *proceedings of the 1st International Workshop on Pattern Recognition and Crowd Analysis, International Conference on Pattern Recognition (ICPR)*. Tsukuba, Japan: [s.n.], 2012. p. 1 – 6.

BAENA-GARCIA, M. et al. Early drift detection method. In: *proceedings of the 4th International Workshop on Knowledge Discovery from Data Streams (ECML/PKDD)*. Berlin, Germany: Springer Berlin Heidelberg, 2006. p. 77 – 86.

BEAUCHEMIN, S. S.; BARRON, J. L. The computation of optical flow. *ACM Computing Surveys*, ACM, New York, USA, v. 27, n. 3, p. 433 – 466, 1995. ISSN 0360-0300.

BERNARD, C. *Wavelets and ill posed problems: optic flow and scattered data interpolation*. Tese (Doutorado) — Laboratory for Applied Mathematics, École Polytechnique, Université Paris, Paris, France, 1999.

BERTINI, M.; BIMBO, A. D.; SEIDENARI, L. Multi-scale and real-time non-parametric approach for anomaly detection and localization. *Computer Vision and Image Understanding*, Elsevier, v. 116, n. 3, p. 320 – 329, 2012. ISSN 1077-3142. Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

BOUZEGZA, M.; ELARBI-BOUDIHIR, M. Automatic understanding of human behavior in videos: A review. In: *proceedings of the 8th International Workshop Systems, Signal Processing and their Applications (WoSSPA)*. Algiers, Algeria: IEEE, 2013. p. 185 – 190.

BRONTE, S. et al. Real-time sequential model-based non-rigid sfm. In: *proceedings of the International Conference on Intelligent Robots and Systems (IROS)*. Illinois, USA: IEEE, 2014. p. 1026 – 1031. Placed by IEEE together with The Robotics Society of Japan (RSJ).

BROX, T.; MALIK, J. Large displacement optical flow: Descriptor matching in variational motion estimation. *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 33, n. 3, p. 500 – 513, 2011. ISSN 0162-8828.

CHANDRASEKARAN, S. et al. An eigenspace update algorithm for image analysis. *Transactions on Graphical Models and Image Processing*, Elsevier, v. 59, n. 5, p. 321 – 332, 1997. ISSN 1524-0703.

CHAQUET, J. M.; CARMONA, E. J.; FERNÁNDEZ-CABALLERO, A. A survey of video datasets for human action and activity recognition. *Transactions on Computer Vision and Image Understanding*, Elsevier, v. 117, n. 6, p. 633 – 659, 2013. ISSN 1077-3142.

CHEN, D. Y.; HUANG, P. C. Motion-based unusual event detection in human crowds. *Transactions on Journal of Visual Communication and Image Representation*, Elsevier, p. 178–186, 2011. ISSN 1095-9076.

CHEN, J. et al. Isomap based on the image euclidean distance. In: *proceedings of the 18th International Conference on Pattern Recognition (ICPR)*. Hong Kong: IEEE, 2006. p. 1110 – 1113.

CHOI, Y. et al. Incremental two-dimensional two-directional principal component analysis (I(2D)2PCA) for face recognition. In: *proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Prague, Czech Republic: IEEE, 2011. p. 1493 – 1496.

CHONG, X. et al. Hierarchical crowd analysis and anomaly detection. *Transactions on Journal of Visual Language and Computing*, Elsevier, v. 25, n. 4, p. 376–393, 2014. ISSN 1045-926X.

CHUNLI, L.; KEJUN, W. A behavior classification based on enhanced gait energy image. In: *proceedings of the 2nd International Conference on Networking and Digital Society (ICNDS)*. Wenzhou, China: IEEE, 2010. v. 2, p. 589 – 592.

COHEN, C. J. et al. Behavior recognition architecture for surveillance applications. In: *proceedings of the 37th Workshop on Applied Imagery Pattern Recognition (AIPR)*. Washington DC, USA: IEEE, 2008. p. 1 – 8.

CONG, Y.; YUAN, J.; LIU, J. Sparse reconstruction cost for abnormal event detection. In: *proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. Rhode Island, USA: IEEE, 2011. p. 3449 – 3456.

CONTE, D. et al. A method for counting moving people in video surveillance videos. *Transactions on European Association for Signal Processing (EURASIP) Journal on Advances in Signal Processing*, Hindawi Pub. Corp. Heidelberg, SpringerOpen, v. 2010, n. 1, p. 1 – 10, 2010. ISSN 1687-6180.

COURTY, N. et al. Using the AGORASET dataset: Assessing for the quality of crowd video analysis methods. *Transactions on Pattern Recognition Letters*, Elsevier, v. 44, p. 161 – 170, 2014. ISSN 0167-8655.

DELAC, K.; GRGIC, M.; LIATSIS, P. Appearance-based statistical methods for face recognition. In: *proceedings of the 47th International Symposium ELMAR, Focused on Multimedia Systems and Applications*. Zadar, Croatia: IEEE, 2005. p. 151 – 158.

DHALL, A.; ASTHANA, A.; GOECKE, R. Facial expression based automatic album creation. In: *proceedings of the 17th International Conference on Neural Information Processing (ICONIP), Models and Applications.* Sydney, Australia: Springer Berlin Heidelberg, 2010. p. 485 – 492.

DU, W. et al. Fuzzy double-threshold track association algorithm using adaptive threshold in distributed multisensor-multitarget tracking systems. In: *proceedings of the International Conference on Green Computing and Communications, Internet of Things and Cyber, Physical and Social Computing (GreenCom-iThings-CPSCom)*. Beijing, China: IEEE, 2013. p. 1133 – 1137.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. New York, USA: Wiley, 2001. v. 2. A Wiley-Interscience Publication.

FAWZY, F.; ABDELWAHAB, M. M.; MIKHAEL, W. 2DHOOF-2DPCA contour based optical flow algorithm for human activity recognition. In: *proceedings of the 56th International Midwest Symposium on Circuits and Systems (MWSCAS)*. New Jersey, USA: IEEE, 2013. p. 1310 – 1313.

FERRYMAN, J.; SHAHROKNI, A. PETS2009: Dataset and challenge. In: *proceedings of the 12th International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*. Utah, USA: IEEE, 2009. p. 1 – 6.

GAMA, J. et al. Learning with drift detection. In: *proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA), Advances in Artificial Intelligence*. Maranhao, Brazil: Springer, 2004. p. 286 – 295.

GATTO, B. B.; HINO, H.; FUKUI, K. Block-based KOMSM for hand shape recognition with occlusion. In: *proceedings of the Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*. Okinawa, Japan: IEEE, 2014. p. 1 – 8.

GU, X.; CUI, J.; ZHU, Q. Abnormal crowd behavior detection by using the particle entropy. *Transactions on International Journal for Light and Electron Optics (Optik)*, Elsevier Ltd., ScienceDirect (distributor), v. 125, n. 14, p. 3428 – 3433, 2014. ISSN 0030-4026.

GUODONG, H. Research on video monitoring system based on intelligent. In: *proceedings of the International Conference on Computer Science and Network Technology (ICCSNT)*. Harbin, China: IEEE, 2011. v. 2, p. 1004 – 1007.

HASSNER, T.; ITCHER, Y.; KLIPER-GROSS, O. Violent flows: Real-time detection of violent crowd behavior. In: *proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Rhode Island, USA: IEEE, 2012. p. 1 – 6.

HSU, S. et al. Falling and slipping detection for pedestrians using a manifold learning approach. In: *proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC)*. Tianjin, China: IEEE, 2013. v. 03, p. 1189 – 1194.

HUEBER, N. et al. Real-time movement detection and analysis for video surveillance applications. In: *proceedings of the Conference on Ground/Air Multisensor Interoperability, Integration and Networking for Persistent ISR V, International Society for Optical Engineering – SPIE*. Maryland, USA: SPIE, 2014. v. 9079, p. 1 – 7.

JIANG, F. et al. Anomalous video event detection using spatiotemporal context. *Transactions on Computer Vision and Image Understanding*, Elsevier, v. 115, n. 3, p. 323 – 333, 2011. ISSN 1077-3142.

JIANG, M. et al. A real-time fall detection system based on hmm and rvm. In: *proceedings of the International Conference on Visual Communications and Image Processing (VCIP)*. Sarawak, Malaysia: IEEE, 2013. p. 1 – 6.

JODOIN, P.; SALIGRAMA, V.; KONRAD, J. Behavior subtraction. *Transactions on Image Processing*, IEEE, v. 21, n. 9, p. 4244 – 4255, 2012. ISSN 1057-7149.

KANEKO, T. et al. A fully connected model for consistent collective activity recognition in videos. *Transactions on Pattern Recognition Letters*, Elsevier, v. 43, p. 109 – 118, 2014. ISSN 0167-8655. ICPR2012 Awarded Papers.

KE, S. et al. A review on video-based human activity recognition. *Transactions on Computers*, MDPI, v. 2, n. 2, p. 88 – 131, 2013. ISSN 2073-431X.

KE, Y.; SUKTHANKAR, R.; HEBERT, M. Event detection in crowded videos. In: *proceedings of the 11th International Conference on Computer Vision (ICCV)*. Rio de Janeiro, Brazil: IEEE, 2007. p. 1 – 8. ISSN 1550-5499.

KIM, B. et al. Compressed sensing with MCT and I(2D)2PCA processing for efficient face recognition. *Transactions on International Journal of Imaging Systems and Technology*, Wiley Periodicals, v. 23, n. 2, p. 133 – 139, 2013. ISSN 1098-1098.

KIM, B.; PARK, H. Efficient face recognition based on MCT and I(2D)2PCA. In: *proceedings of the International Conference on Systems, Man, and Cybernetics (SMC)*. Seoul, South Korea: IEEE, 2012. p. 2585 – 2590.

KIM, S.; MALLIPEDDI, R.; LEE, M. Incremental face recognition using rehearsal and recall processes. In: *proceedings of the International Joint Conference on Neural Networks (IJCNN)*. Beijing, China: IEEE, 2014. p. 2752 – 2757.

KRAUSZ, B.; BAUCKHAGE, C. Loveparade 2010: Automatic video analysis of a crowd disaster. *Transactions on Computer Vision and Image Understanding*, Elsevier, v. 116, n. 3, p. 307 – 319, 2012. ISSN 1077-3142.

LI, T. et al. Crowded scene analysis: A survey. *Transactions on Circuits and Systems for Video Technology*, IEEE, v. 25, n. 3, p. 367 – 386, 2014. ISSN 1051-8215.

LIU, J. et al. An algorithm of auto-update threshold for singularity analysis of pipeline pressure. *Transactions on Mathematical Problems in Engineering: Theory, Methods, and Applications*, Hindawi Publishing Corporation, v. 2013, n. 495425, 2013. ISSN 1024-123X.

LIU, Y.; LI, X.; JIA, L. Abnormal crowd behavior detection based on optical flow and dynamic threshold. In: *proceedings of the 11th World Congress on Intelligent Control and Automation (WCICA)*. Shenyang, China: IEEE, 2014. p. 2902 – 2906.

PATEL, A.; NGUYEN, T.; BARANIUK, R. G. A probabilistic theory of deep learning. *ArXiv e-prints*, SAO/NASA Astrophysics Data System, v. 1, n. 1, p. 1 – 56, 2015. ArXiv:1504.00641v1, Technical Report Number 2015-1.

PERŠ, J. et al. Histograms of optical flow for efficient representation of body motion. *Transactions on Pattern Recognition Letters*, Elsevier, v. 31, n. 11, p. 1369 – 1376, 2010. ISSN 0167-8655.

POPOOLA, O. P.; WANG, K. Video-based abnormal human behavior recognition - a review. *Transactions on Systems, Man, and Cybernetics (Applications and Reviews), Part C*, IEEE, v. 42, n. 6, p. 865 – 878, 2012. ISSN 1094-6977.

RASHEED, N.; KHAN, S. A.; KHALID, A. Tracking and abnormal behavior detection in video surveillance using optical flow and neural networks. In: *proceedings of the 28th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. British Columbia, Canada: IEEE, 2014. p. 61 – 66.

RODRÍGUEZ, N. D. et al. A survey on ontologies for human behavior recognition. *Transactions on Computing Surveys (CSUR)*, ACM, v. 46, n. 4, p. 1 – 33, 2014. ISSN 0360-0300. Article No. 43.

ROSHTKHARI, M. J.; LEVINE, M. D. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Transactions on Computer Vision and Image Understanding*, Elsevier, v. 117, n. 10, p. 1436 – 1452, 2013. ISSN 1077-3142.

ROTH, P. M.; WINTER, M. Survey of appearance-based methods for object recognition. *Technical Report ICG-TR-01/08*, Institute for Computer Graphics and Vision, Graz University of Technology, Styria, Austria, p. 1 – 68, 2008.

SAINI, M. et al. Adaptive workload equalization in multi-camera surveillance systems. *Transactions on Multimedia*, IEEE, v. 14, n. 3, p. 555 – 562, 2012. ISSN 1520-9210.

SAINI, M. K.; ATREY, P. K.; SADDIK, A. E. From smart camera to smarthub: Embracing cloud for video surveillance. *Transactions on International Journal of Distributed Sensor Networks*, Hindawi Publishing Corporation, v. 2014, n. 757845, p. 1 – 10, 2014. ISSN 1550-1329.

SANGUANSAT, P. Two-dimensional principal component analysis and its extensions. *Transactions on Principal Component Analysis, InTech Open Access Books*, INTECH Open Access Publisher, n. 2005, p. 1 – 23, 2010. DOI: 10.5772/36892.

SEO, J.; KIM, S. D. Recursive on-line (2D)2PCA and its application to long-term background subtraction. *Transactions on Multimedia*, IEEE, v. 16, n. 8, p. 2333 – 2344, 2014. ISSN 1520-9210.

SINDHUJA, C. R.; SRINIVASAGAN, K. G.; KALAISELVI, S. An efficient method for crowd event recognition based on motion patterns. In: *proceedings of the International Conference on Recent Trends in Information Technology (ICRTIT)*. Tamil Nadu, India: IEEE, 2014. p. 1 – 6.

VISHWAKARMA, S.; AGRAWAL, A. A survey on activity recognition and behavior understanding in video surveillance. *Transactions on The Visual Computer*, Springer, v. 29, n. 10, p. 983 – 1009, 2012. ISSN 0178-2789.

WALHA, A.; WALI, A.; ALIMI, A. M. A system of abnormal behaviour detection in aerial surveillance. In: *proceedings of the 9th International Conference on Information Assurance and Security (IAS)*. Tunis Governorate, Tunisia: IEEE, 2013. p. 102 – 107.

WANG, L.; ZHANG, Y.; FENG, J. On the euclidean distance of images. *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 27, n. 8, p. 1334 – 1339, 2005. ISSN 0162-8828.

WANG, Z. et al. Image quality assessment: from error visibility to structural similarity. *Transactions on Image Processing*, IEEE, v. 13, n. 4, p. 600 – 612, 2004. ISSN 1057-7149.

YANG, J. et al. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 26, n. 1, p. 131 – 137, 2004. ISSN 0162-8828.

YILMAZ, A.; JAVED, O.; SHAH, M. Object tracking: A survey. *Transactions on Computing Surveys (CSUR)*, ACM, v. 38, n. 4, p. 1 – 45, 2006. ISSN 0360-0300. Article No. 13.

ZHAO, F.; LI, J. Pedestrian motion tracking and crowd abnormal behavior detection based on intelligent video surveillance. *Transactions on Journal of Networks*, Academy Publisher, v. 9, n. 10, p. 2598 – 2605, 2014. ISSN 1796-2056.

ZHAO, H.; YUEN, P. C.; KWOK, J. T. A novel incremental principal component analysis and its application for face recognition. *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, IEEE, v. 36, n. 4, p. 873 – 886, 2006. ISSN 1083-4419.

ZIN, T. T. et al. Unattended object intelligent analyzer for consumer video surveillance. *Transactions on Consumer Electronics*, IEEE, v. 57, n. 2, p. 549 – 557, 2011. ISSN 0098-3063.

ZIN, T. T. et al. An integrated framework for detecting suspicious behaviors in video surveillance. In: *proceedings of the Conference on Video Surveillance and Transportation Imaging Applications, International Society for Optical Engineering – SPIE*. California, USA: SPIE, 2014. v. 9026, p. 1 – 8.