Detecção de Anomalias Em Vídeos Utilizando Dicionários Espaço-Temporais

Mateus T. Nakahata, Eduardo A. B. da Silva e Sergio L. Netto

***Resumo*— Este trabalho apresenta uma implementação do método *Spatio-Temporal Compositions* (STC) para a detecção de anomalias em vídeo. O STC, assim como o *Bag of Video words* (BOV), utiliza um dicionário para eliminar as redundâncias, porém difere deste por levar em consideração a disposição espaço-temporal de pequenos volumes de vídeos. Além disso, o STC realiza uma modelagem utilizando uma abordagem probabilística, na qual eventos anômalos são aqueles com baixa probabilidade de ocorrência. Modificações incorporadas ao STC incluem o grau de superposição dos volumes, etapa de filtragem espaço-temporal e cálculo da probabilidade em cada escala. O algoritmo utilizado neste modelo apresenta bons resultados na identificação das anomalias, sem a subtração de plano de fundo, estimação de movimento ou rastreamento. O sistema é preciso mesmo com um pequeno treinamento e sem o conhecimento prévio do tipo de evento a ser observado, sendo robusto a variações de luminosidade e grau de complexidade do ambiente em questão, conforme ilustrado por diversos exemplos.**

***Palavras-Chave*— Detecção de anomalias, pacote de palavras de vídeo, composições espaço-temporais, vídeo vigilância.**

Abstract*— *This paper presents an implementation of the Spatio-Temporal Compositions (STC) method for the detection of video anomalies. The STC, as the Bag of Video words (BOV), uses a dictionary to eliminate to redundancies, but differs from it by taking into consideration the spatio-temporal composition of small volumes of videos. The STC also performs a modeling using a probabilistic approach, in which anomalous events are those with a low probability of occurrence. Some modifications incorporated to the STC include volume superposition level, spatio-temporal filtering stage and multi-scale computation of probability function. The algorithm used in this model gives good results in the identification of anomalies, without background subtraction, motion estimation or tracking. The system is accurate even with a little training and no prior knowledge of the type of event to be observed, being robust to light variations or cluttered environments, as illustrated by several examples.

Keywords*— *Anomalies detection, bag of video words, spatio-temporal compositions, video surveillance.

I. Introdução

Os sistemas de videovigilância são cada vez mais utilizados na segurança pública e privada [1]. Além disso, por questões de segurança nas indústrias, são utilizados para que os operadores trabalhem em um ambiente seguro e confortável. No entanto, os operadores são expostos a um grande volume de imagens e em regime de 24x7, o que torna a vigilância sujeita a erros. No intuito de minimizar este problema, sistemas automáticos de análise de vídeo são cada vez mais utilizados. Porém estes sistemas precisam ser treinados e configurados para o seu correto funcionamento. Isso requer um conhecimento prévio dos eventos de interesse, o que nem sempre é possível. A busca é por anomalias ou situações não comuns, que representam ameaças. Além disso, muitas vezes os ambientes monitorados são tumultuados ou mudam durante o passar do tempo, e os sistemas de análise de vídeo necessitam ser constantemente reconfigurados.

No método *Bag of Video words* (BOV) a análise é realizada através de pequenos volumes espaço-temporais, e a redundância entre os mesmos é minimizada através da utilização de um dicionário [2]. Estes métodos apresentam um melhor desempenho em ambientes desordenados, porém a interpretação da imagem pelo ser humano é influenciada pela disposição espaço-temporal entre os objetos que compõem esta imagem [3], o que não é considerado no BOV. O método *Spatio-Temporal Composition* (STC) [4] leva em consideração a disposição espaço-temporal de pequenos volumes de vídeos e realiza uma modelagem utilizando uma abordagem probabilística, na qual eventos anômalos são aqueles com baixa probabilidade de ocorrência. Além disso, o STC possui como características a possibilidade de utilização em tempo real bem como o contínuo autotreinamento, sendo capaz de adaptar-se conforme as condições ambientais mudam. Este método também não necessita de configurações prévias para a detecção das anomalias.

Este trabalho traz um pequeno resumo do método STC, e busca complementar o trabalho de [4], descrevendo as etapas a serem realizadas e algumas modificações por nós incorporadas. Exemplos de modificações incluem o tamanho da superposição dos volumes, a influência dos tamanhos das células, necessidade de filtragem espaço-temporal para minimizar os ruídos e forma do cálculo da probabilidade com várias escalas. Por fim, é realizada uma avaliação da qualidade dos resultados obtidos. Os testes da implementação foram realizados com vários vídeos, primeiro treinando o programa com uma pequena amostra de uma cena considerada normal, e em seguida analisando um vídeo similar ao de teste mas com uma pequena diferença na cena.

Para a programação, utilizou-se o QT [5], que é um programa de desenvolvimento em C++, multiplataforma e gratuito para fins não comerciais, bem como a biblioteca Opencv [6], que é uma biblioteca de código aberto com funções que implementam algoritmos de visão computacional.

As imagens utilizadas nos testes foram da biblioteca de imagens *UCSD Anomaly Detection Dataset* [7]. Estas imagens foram adquiridas de uma câmera montada em uma posição elevada, gravando a imagem de uma área ampla onde diver-

Mateus T. Nakahata, Eduardo A. B. da Silva e Sergio L. Netto Departament of Electrical Engineering, Federal University of Rio de Janeiro, Rio de Janeiro-RJ, Brazil, E-mails: mateus.nakahata@smt.ufrj.br, eduardo@smt.ufrj.br, sergioln@smt.ufrj.br.

sas pessoas transitam, em sua maioria caminhando. Também foram realizados testes com outra base de dados denominada CAVIAR [8], na qual tem-se uma câmera de vigilância em um ponto elevado de um saguão.

II. MÉTODO STC (*Spatio-Temporal Composition*)

Neste método, novas amostras de vídeo são decompostas em pequenos volumes representados por palavras de um dicionário. Em seguida, são calculadas as probabilidades de ocorrência das composições espaço-temporais formadas por essas palavras. Composições com baixa probabilidade são candidatas a serem anômalas. Nesta seção, serão discutidas resumidamente o treinamento e análise dos vídeos, conforme o trabalho de Roshtkhari e Levine [4]. A figura 1 descreve os passos principais do método para identificar as anomalias nas imagens. O treinamento é realizado com uma pequena amostra de vídeo com uma cena considerada normal. As etapas iniciais de amostragem e criação do descritor são idênticas no treinamento e análise. Na figura 1 as etapas em cinza são aquelas que necessitaram de um estudo mais detalhado sobre a sua forma de implementação e nas quais algumas mudanças foram incorporadas.

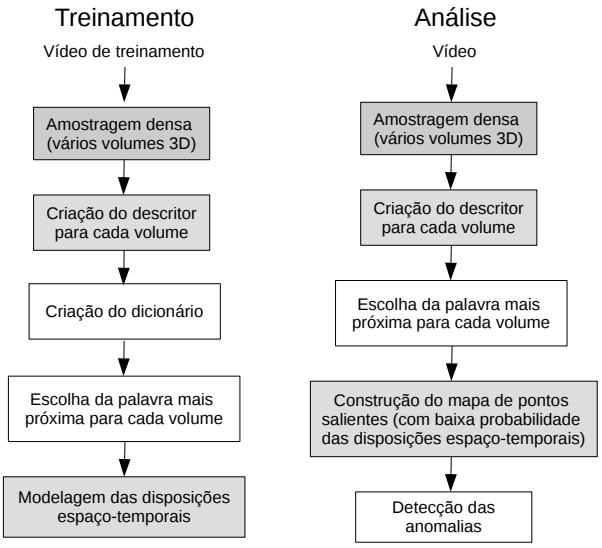

Fig. 1. Sequência das etapas de treinamento e análise de vídeo. As etapas em cinza tiveram uma implementação realizada de forma modificada neste trabalho.

A. Amostragem e Criação do Descritor

A amostragem do conteúdo do vídeo é baseada em *Bag of Video words* (BOV), que consiste em volumes espaço-temporais obtidos através de amostragem densa, que procura manter as informações relevantes do vídeo [9].

Um dicionário é criado com o intuito de reduzir a redundância entre os volumes de vídeo. Para isso, o vídeo é dividido em pequenos volumes 3D, $p_i \in \Re^{n_x \times n_y \times n_t}$, em torno de cada *pixel*, onde $n_x \times n_y$ é uma pequena área do quadro e n_t representa uma pequena variação no tempo. Esta decomposição espaço-temporal é realizada em várias escalas no espaço e no tempo, gerando uma pirâmide de segmentos de vídeo.

Cada volume p_i é representado por um descritor g_i que é simplesmente o valor absoluto da derivada temporal $\triangle_t$ de cada *pixel* do volume p_i, conforme

$$\forall p_i, g_i = abs(\triangle_t(p_i)). \quad (1)$$

Os valores obtidos para cada *pixel* de p_i são empilhados em um vetor e normalizados como um valor unitário, criando um descritor compacto em várias escalas. Este descritor é robusto mesmo para ambientes onde o plano de fundo não é estático e apresenta apenas pequenas variações. Outros descritores podem apresentar melhores resultados, dependendo da aplicação, como por exemplo os utilizados em [10], [11].

Para compor o dicionário, primeiro passo foi a filtragem das amostras de vídeo, onde é realizada uma varredura em todos os pixels da imagem, onde são escolhidas oito amostras, conforme a figura 2 (a). Os círculos representam o centro de cada pixel, que formam um cubo. Em seguida é calculada a média destes pixels, e formada uma nova imagem com estas médias.

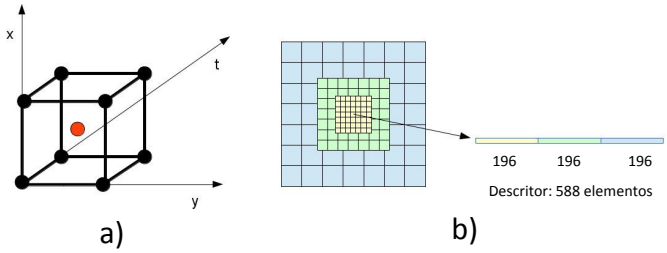

Fig. 2. a) A filtragem é realizada calculando a média de oito pixels, quatro em um quadro e quatro no quadro seguinte. b) Construção do descritor, concatenando as amostras das derivadas em várias escalas.

Para a construção de descritor em várias escalas, primeiro construímos o descritor da maior escala, vetorizando o volume 7x7x4 pixels em um vetor de 196 pixels. O tamanho do volume em pixels foi definido empiricamente, de acordo com [4]. Em seguida calculamos em qual ponto o pixel central está localizado nas demais escalas. Após a identificação deste ponto central, é realizada a vetorização dos 7x7x4 pontos centralizados neste ponto central. Os vetores em cada escala são concatenados formando um novo descritor, de tamanho 588, conforme a figura 2 (b).

Uma questão a ser respondida é se ocorre ou não a superposição dos volumes. Em [11] uma superposição de 50% é utilizada com resultados satisfatórios, obtendo um compromisso entre precisão e tempo de processamento. Nos testes realizados, quanto maior a superposição melhores os resultados, até o limite de se realizar a amostragem pixel a pixel, porém com grande aumento no tempo de processamento, como era de se esperar. Realizou-se então uma sobreposição espacial de 50%, ou 3 pixels, pois com uma superposição maior o tempo de processamento aumenta consideravelmente,de forma geometrica. No tempo, o espaçamento depende da taxa de quadros por segundo utilizada. Com uma taxa de 5 fps (*frames per second*), foi feita uma amostragem quadro a quadro.

B. Criação do Dicionário e Escolha da Palavra Para Cada Volume

Devido à amostragem densa e às várias escalas utilizadas, o número de volumes espaço-temporais é muito grande, e como estes volumes possuem muita redundância entre si, os volumes similares são agrupados, e para cada grupo é criado um descritor que é salvo no dicionário. O dicionário é criado utilizando-se métodos de agrupamento, como por exemplo *k-means*. A etapa de criação do dicionário foi feita através da implementação direta do algoritmo descrito. O único parâmetro a ser configurado foi o número máximo de palavras no dicionário, ajustado para 20, já que valores superiores a este acarretam uma melhoria insignificante, conforme os testes realizados por [4]. A título de comparação, foi feito o teste da criação dos códigos através de uma mistura de gaussianas, com 20 gaussianas, ao invés de utilizar o código proposto. Não foi observado nenhuma melhoria significativa nos resultados de objetos detectados e falso positivos, e o tempo de processamento aumentou muito,na ordem de dezenas de vezes maior. O código original possui como característica se adaptar a cada nova amostra, o que é útil no caso onde o treinamento é contínuo. Em ambiente externo, por exemplo, o algoritmo pode se adaptar à medida que as condições de luz solar se alteram, quadro a quadro, tornando a atualização do código gradual.

Após a criação do dicionário, um código foi alocado a cada volume da imagem de treinamento. O critério para esta alocação foi a menor distância euclidiana. Após a criação do dicionário, cada volume v_i é relacionado com a palavra c_j com um peso $w_{i,j}$ dado por

$$w_{i,j} = \frac{1}{\Sigma_j \frac{1}{distância(v_i,c_j)}} \times \frac{1}{distância(v_i,c_j)}. \quad (2)$$

Isto é feito tanto na etapa de treinamento quanto na etapa de detecção de anomalias, e mede o grau de proximidade do descritor do volume para cada palavra do dicionário.

C. Modelagem das Disposições Espaço- Temporais

A maioria dos métodos que utilizam BOV não leva em consideração o arranjo espaço-temporal entre os volumes ou limita-se a um pequeno volume ao redor do ponto de amostragem. Neste método é utilizada uma abordagem probabilística para determinar se o volume é anômalo ou não, baseado na probabilidade do arranjo dos volumes dentro de uma região maior.

A representação do conjunto é feita da seguinte forma: seja E_i o conjunto centralizado no ponto (x_i, y_i, t_i) em coordenadas absolutas e contendo K volumes. Utilizam-se as coordenadas relativas para determinar a posição dos volumes dentro do conjunto, conforme a figura 3 (a). Dado o volume v_k dentro do conjunto E_i, define-se $\Delta_{v_k}^{E_i} \in \Re^3$ como a posição relativa (no espaço e no tempo) de v_k, localizado no ponto (x_k, y_k, t_k), dentro de E_i:

$$\Delta_{v_k}^{E_i} = (x_k - x_i, y_k - y_i, t_k - t_i). \quad (3)$$

Desta forma, o conjunto E_i de volumes, centrado na posição (x_i, y_i, t_i), é inicialmente representado como um conjunto de volumes de vídeo e suas posições relativas em relação ao volume central:

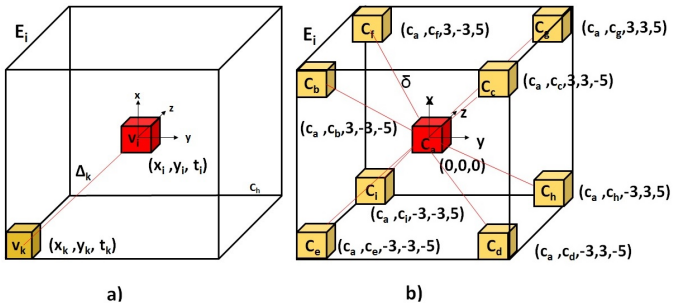

Fig. 3. Posição relativa dos volumes dentro do conjunto. O conjunto é representado pelo arranjo espaço-temporal das palavras, que estão a uma distância δ da palavra central c'.

$$E_i = \{\Delta_{v_k}^{E_i}, v_k, v_i\}_{k=1}^{K}. \quad (4)$$

Cada volume v_k do conjunto é vinculado com a palavra $c_j \in \mathbf{C}$ com um peso w_j, que representa a sua similaridade. Sendo assim, o arranjo dos volumes pode ser representando por um conjunto de palavras e seu arranjo espaço-temporal. Seja $\nu \subset \Re^{n_x \times n_y \times n_t}$ o espaço dos descritores de um volume de vídeo, e $\mathbf{C}$ seu dicionário; $c : \nu \rightarrow \mathbf{C}$ é uma variável aleatória alocando uma palavra a um volume de vídeo e $c' : \nu \rightarrow \mathbf{C}$ é uma variável aleatória designando uma palavra para o volume central do conjunto. Desta forma $\delta : \Re^3 \rightarrow \Re^3$ é uma variável aleatória representando a distância relativa da palavra do ponto central até a palavra c. Portanto pode-se representar o conjunto de volumes como um arranjo de palavras do dicionário, conforme a figura 3 (b). Ou seja, ao invés de representar o conjunto E_i como um arranjo de volumes, representa-se como um arranjo de palavras.

Sejam $O = (v_k, v_i, \Delta_{v_k}^{E_i})$ a observação do volume v_k em relação ao volume central v_i dentro do conjunto E_i, e $\Delta_{v_k}^{E_i}$ a posição relativa do volume observado v_k com relação à v_i dentro de E_i. O objetivo é medir a probabilidade de cada hipótese $h = (c, c', \delta)$, obtida pela substituição dos volumes por palavras do dicionário, dada a observação O,

$$P(h/O) = P(c, c', \delta \mid v_k, v_i, \Delta_{v_k}^{E_i}). \quad (5)$$

Pode-se demonstrar que [4]:

$$P(c, c', \delta \mid v_k, v_i, \Delta_{v_k}^{E_i}) = P(\delta \mid v_k, v_i, \Delta_{v_k}^{E_i}) P(c' \mid v_i) P(c \mid v_k). \quad (6)$$

Ou seja, em um conjunto ao redor do pixel, com um volume central v_i, e outros volumes v_k dentro deste conjunto a uma distância $\Delta_{v_k}^{E_i}$ do volume central, quer-se calcular a probabilidade de se atribuir a palavra c' ao volume central e c aos demais volumes. A probabilidade $P(\delta \mid v_k, v_i, \Delta_{v_k}^{E_i})$ é determinada através da aproximação da sua *pdf* por uma mistura de gaussianas, utilizando *expectation maximization* [12], sendo que as amostras são os arranjos observados anteriormente, durante o treinamento. Portanto, as amostras passadas formam um vetor de amostra $\mathbf{a}(c_i, c_k, \delta)$, onde δ é a distância relativa entre as palavras. Várias destas amostras permitem estimar a *pdf*, conforme a figura 4. As probabilidades $P(c' \mid v_i)$ e

$P(c \mid v_k)$ de cada volume espaço-temporal é calculada durante a alocação das palavras.

A probabilidade a posteriori é calculada de acordo com

$$P(c_j \mid v_i) = \frac{w_{i,j} \times P(c_j)}{\sum_j w_{i,j} \times P(c_j)}. \quad (7)$$

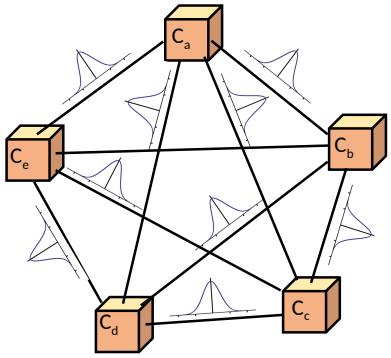

Fig. 4. É calculada a pdf do arranjo 3D dos volumes v_k 2 a 2 com palavras associadas c_k dentro de cada conjunto E_i. O dicionário é formado pelas palavras em conjunto com a pdf.

Nas simulações realizadas, utilizou-se um conjunto de 7x7x11 volumes. A amostragem é realizada da seguinte forma: para cada volume com descritor c_0 dentro da imagem de treinamento, são consideradas as posições relativas dos 539 volumes ao redor de c_0, estabelecendo-se uma conexão entre estes volumes. Cada conexão é representada por um vetor $\mathbf{v}(x, y, t, c_0, c_i)$ em $\Re^5$.

D. Detecção de Padrões Anômalos e Construção do Mapa de Pontos Salientes

Na Análise a etapa de decomposição da imagem em volumes e vetorização são iguais às do treinamento. Em seguida é medida a distância entre o volume e cada palavra do dicionário pela equação (2).

A equação (6) representa a probabilidade de atribuição da palavra para apenas uma das relações entre o volume central e os demais K volumes do conjunto E_i.

Conforme demonstrado em [4], a probabilidade máxima a posteriori de todos os volumes v_k dentro de E_i centrados em v_i pode ser escrita como:

$$\max_{\substack{c \in C \\ c' \in C}} P(c, c', \delta \mid E_i^Q) = \max_{\substack{c \in C \\ c' \in C}} \prod_k^K P(\delta \mid c, c', \Delta_{q_k}^{E_i^Q}) P(c \mid q_k) P(c' \mid q_i). \quad (8)$$

De forma resumida, o vídeo a ser analisado Q, ou *query*, é amostrado densamente em várias escalas espaço-temporais, construindo os volumes de vídeo v. Para cada v_k é alocada uma palavra $c \in C$ com uma similaridade w. A probabilidade de cada pixel ser normal ou uma anomalia é calculada considerando-se o arranjo espaço-temporal dos volumes dentro do conjunto E_i^Q.

Dado que em um conjunto E_i têm-se K volumes, e que o dicionário possui M palavras, primeiramente atribui-se ao volume central c_i a primeira palavra do dicionário. Em seguida, em todos os volumes c_k são testados todas as M palavras e escolhida a que maximiza a probabilidade conforme (8). Repete-se o processo testando-se c_i com todas as palavras. A escolha será a atribuição de palavras que maximiza a probabilidade de E_i. Sendo assim, a ordem de grandeza de operações necessárias para o cálculo da probabilidade de E_i é de $O(K \times M \times M)$. Portanto, é importante manter o número de palavras baixo, caso contrário o processo torna-se lento.

Para cada bloco é calculada a sua probabilidade de ocorrência. Blocos com baixa probabilidade de ocorrência são marcados em vermelho, gerando manchas nas regiões de baixa probabilidade. Para eliminar os ruídos, foi feita uma pós-filtragem com um filtro passa-baixa, eliminando pontos isolados, que geralmente aparecem piscando na imagem. Além disso, foi feita uma dilatação nas manchas, para obter-se um contorno mais suave e juntar manchas muito próximas.

Por fim, é realizada uma votação da seguinte forma: para uma mancha ser considerada como uma região de interesse, é necessário que a mesma esteja presente em pelo menos oito de cada dez quadros. Isso evita que manchas geradas por ruídos sejam consideradas. Considera-se que sejam necessários pelo menos oito para evitar o caso oposto, no qual devido a ruídos uma mancha de interesse possa ser desconsiderada. Para esta votação, é necessário identificar as manchas em cada quadro e em seguida estimar o seu deslocamento no quadro seguinte.

III. Resutados Obtidos

Para os testes, o conceito de anomalia utilizado foi o de eventos que diferem muito dos observados no vídeo de teste. O critério é subjetivo, e depende do limiar de probabilidade utilizado. Inicialmente realizou-se um treinamento com o vídeo curto de cerca de dez segundos onde apenas estão presentes pessoas andando, e os resultados obtidos na detecção de anomalias no primeiro vídeo de teste são mostrados na figura 5, onde apenas o ciclista é detectado. Na figura 6 a diferença

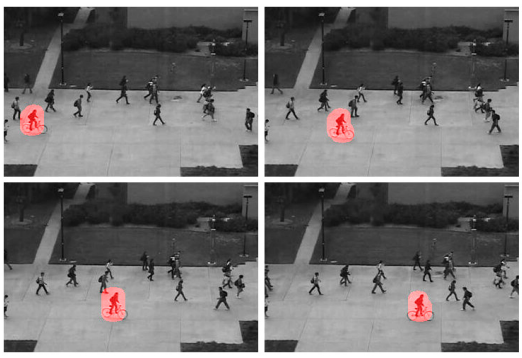

Fig. 5. Apenas o ciclista é detectado. As pessoas andando não são detectadas pois no vídeo de treinamento existiam várias pessoas andando de forma parecida.

mais significativa são um carrinho e um ciclista, trafegando da direita para a esquerda. Ambos os exemplos os objetos foram detectados adequadamente, e as pessoas andando não foram

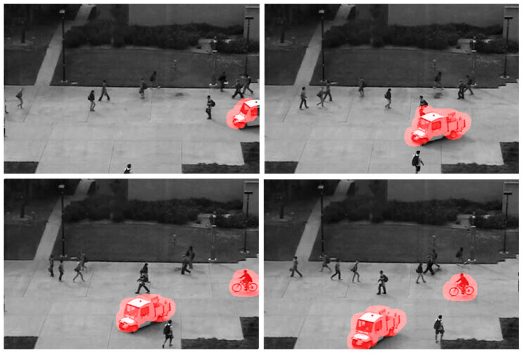

Fig. 6. Tanto o carrinho quanto o ciclista foram reconhecidos como objetos de interesse. As pessoas andando não são detectadas.

consideradas como anomalias. Foram realizados outros testes com diversos vídeos da mesma biblioteca, e os resultados foram similares.

Também foram realizados testes com outra biblioteca, com uma câmera em um saguão. Na figura 7, duas pessoas se encontram no meio do saguão. Na figura 8, as pessoas

Fig. 7. Duas pessoas se encontram no meio do saguão e depois seguem juntas. O evento é detectado como uma anomalia pois não há cena parecida no vídeo de treinamento.

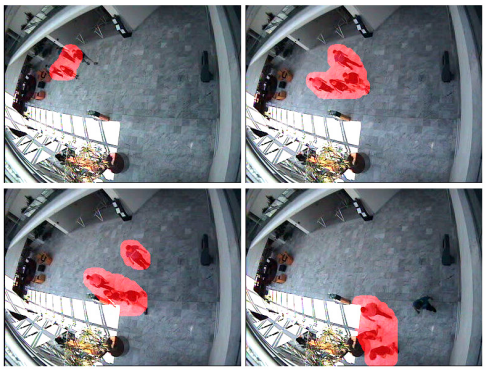

Fig. 8. Diversas pessoas atravessam o saguão e são detectadas.A pessoa mais a esquerda não é detectada na parte final do video, pois no vídeo de treinamento há uma pessoa realizando este trajeto, do meio do saguão para a saída.

atravessando o saguão são detectadas como anomalias, pois não há evento similar no treinamento. Porém a pessoa mais à direita do vídeo não é marcada na parte final do vídeo, pois no treinamento uma pessoa realiza um trajeto similar, apenas nesta parte do saguão.

IV. Conclusões

Neste trabalho analisou-se o método STC para detecção de anomalias em vídeo, detalhando e desenvolvendo as etapas que não apresentam uma explicação clara no artigo de referência. Observa-se que o método STC não visa o reconhecimento de eventos já conhecidos, e sim identificar eventos estranhos, sendo assim um complemento aos métodos tradicionais, não um substituto. Os resultados obtidos foram os esperados na identificação das anomalias, sem a utilização de subtração de plano de fundo, estimação de movimento ou rastreamento, mesmo com um pequeno treinamento e sem o conhecimento prévio do tipo de evento a ser observado. Também foi possível identificar eventos em ambientes confusos ou tumultuados. Foram propostas etapas complementares como filtragem passa-baixa, tamanho da amostragem dos volumes, pós-processamento com a dilatacão e votação, que se mostraram muito importantes para se obter bons resultados. Com os resultados obtidos foi possível adquirir um conhecimento dos principais parâmetros que influenciam o método, que servirá de subsídio para a futuros trabalhos, como a adaptação do mesmo para a utilização em uma câmera em movimento.

O código fonte desenvolvido em C++ e um manual de utilização encontram-se disponíveis no endereço http://www.smt.ufrj.br/ eduardo/stc/ .

Referências

[1] Haering, N.,Venetianer, P. L., Lipton, A. *The evolution of video surveillance: An overview*, In: Machine Vision and Applications, vol.19, no. 5-6, pp. 279-290, June 2008.

[2] Lazebnik, S., Schmid, C., Ponce, J. *Beyond bags of features: spatial pyramid matching for recognizing natural scene categories*, In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169-2178, New York, June 2008.

[3] Schwartz, O., Hsu, A., Dayan, P. *Space and time in visual context*, In: Nature Reviews Neuroscience, vol.8, no. 7, pp. 522-535, July 2007.

[4] Roshtkhari, M. J., Levine, M. D. *An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions*, In: Computer Vision and Image Understanding, vol.117, no. 10, pp. 1436-1452, July 2013.

[5] The QT Company *QT Project*, http://www.qt-project.org, acessado em 10 de Abril de 2015.

[6] OPENCV *(Open Source Computer Vision Library*, http://www.opencv.org, acessado em 10 de Abril de 2015.

[7] Li, W., Mahadevan, V., Vasconcelos, N. *UCSD Anomaly Detection Dataset*, http://www.svcl.ucsd.edu/projects/anomaly, acessado em 10 de Abril 2015.

[8] Fisher, R. *Caviar Project*, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/, acessado em 10 de Abril 2015.

[9] Rapantzikos, K., Avrithis, Y., Kollias, S. *Dense saliency-based spatio-temporal feature points for action recognition*, In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2009, pp. 1454-1461, Miami Beach, June 2009.

[10] Zhong, H., Shi, J., Visontai, H. *Detecting unusual activity in video*, In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2004, pp. 819-826, Washington, June 2004.

[11] Bertini, M., Del Bimbo, A., Seidenari, L. *Multi-scale and real-time non-parametric approach for anomaly detection and localization*, In:Computer Vision and Image Understanding, vol.116, no. 3, pp. 320-829, March 2012.

[12] Bilmes, J. *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, International Computer Science Institute and Computer Science Division, University of California, Berkeley, 1998.