Abnormal Motion Analysis for Tracking-Based Approaches Using Region-Based Method with Mobile Grid

Jorge Henrique Busatto Casagrande
Instituto Federal de Santa Catarina–IFSC/Núcleo de Telecomunicações, São José, Santa Catarina, Brasil
Email: casagrande@ifsc.edu.br

Marcelo Ricardo Stemmer
Universidade Federal de Santa Catarina–UFSC/Departamento de Automação e Sistemas, Florianópolis, Santa Catarina, Brasil
Email: marcelo.stemmer@ufsc.edu.br

***Abstract*—Tracking-based video surveillance approaches use a pipe line of processes from capture of frames up to video analysis. All these processes consume too much computational cost and generally it is concentrated in the last step of this framework. Particularly for this step, our paper proposes a method for abnormal motion analysis that ensures efficiency in the inferences with less computational effort. For this, we use a region-based model that uses a mobile grid of subregions constructed from scene's ROI (region of interest). In order to avoid the implementation of the complete framework, we have replaced the previous steps with annotated datasets from the real world. From these annotations, we seek a size of subregion that produces the best result in the abnormal motion detection using GMM (Gaussian Mixture Models) and ROC (Receiver Operating Characteristic) curves. The method proved efficient and useful for abnormal motion analysis, especially in tracking-based approaches.**

***Index Terms*—motion analysis, abnormal motion detection, pattern recognition, video surveillance**

I. Introduction

Automated surveillance has received much attention in recent years especially to support the tedious work of those who operate traditional systems. In this aspect, the motion analysis is one of the most explored research lines nowadays [1]. The region-based or clustering-based approaches using spatio-temporal probability models are appearing as the most effective approaches for motion analysis. The inherent uncertainty of the observations in video scenes is a characteristic problem, which reinforces the use of probabilistic reasoning in the events modeling. Some proposals had to determine constraints on their models in order to reduce the computation workload involved in every process part [2]-[5]. Several authors developed complete video surveillance systems, from the capture of video frames up to the behavior analysis of moving objects in a category called tracking-based, where the robust tracking of multiple objects is still an open problem. Fig. 1 shows a taxonomy of the main lines of research in the abnormal motion analysis. Our work is in the branches outlined with an ellipse.

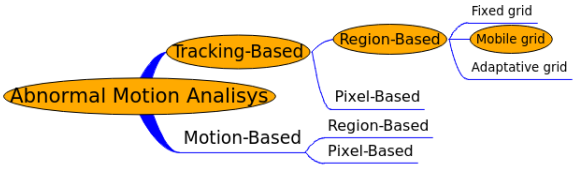

Figure 1. Abnormal motion analysis taxonomy.

The research seeks strategies that require lower computational cost, in order to make applications involving real-world scenarios feasible in different contexts [6]. Several papers use a motion-based approach for the abnormal motion analysis [7]-[14]. This category is attractive because it requires no preprocessing of video. The motion-based proposals, advance to other fields of research such as crowd, gait, gesture, face analysis among others.

Especially in tracking-based approaches, the large data amount required to be processed and the algorithm's complexity are considered as a barrier for the computational treatment. In that sense, many approaches deal with real-world scenes, but are generally limited in flexibility in what concerns scenarios, targets, video length and reality. A usual way to work around the problems of computational overhead is the adoption of models region-based where the analysis is done on clusters of pixels. Authors such as Elhoseiny [15] agree that it is necessary to use region-based techniques, otherwise it is impracticable to apply their ideas in the real world. In recent paper [16], we have presented a

Manuscript received January 27, 2014; revised May 5, 2014.

doi: 10.12720/joig.2.1.22-27

region-based abnormal motion detection using a fixed grid. This work has shown very good results, which ensured a best margin of correct abnormal motions detection for each type of scenario, even with a significant reduction of data samples. Our goal here is to keep the same proposal however using a mobile grid rather than a fixed grid.

A. Proposed Approach

In a tracking-based approach the abnormal motion analysis is the last step of a framework. Thus, to take faster our goal, we isolate this step and have replaced the previous steps with video annotations. This strategy avoids the development of all the involved processes since the capture of video frames, which are not the purpose of this work. We implemented our abnormal motion analysis model which takes as reference input, data annotated from video dataset. In one round of the our training model, the mobile grid is formed, the motion model is performed and finally we have the best decision threshold to be used in respective scenario until that new round will be necessary. Since known the best threshold, it can be used for test step or any video of same scene while is kept the same behavior and frequency of the mobile objects. For motion and learning modeling, we adopted similar models proposed by the authors Basharat *et al.* [2], however we use a region-based approach instead pixel-based. In our learning model, beyond the use of GMM trained by the EM algorithm, we adopted a supervised and off-line training model and single-class classification to simplify the implementation.

As a reference dataset, we used two sets containing 8 to 10 sequences of 30 minutes of the LOST Project videos (Longterm Observation of Scenes with Tracks Dataset) available by the authors Abrams et al. [17] in *http://lost.cse.wustl.edu*. The LOST dataset comprises several videos made from *streaming* of outdoor *webcams*, captured and organized by numbers (1 to 25) in the same half hour every day at various locations around the world. The dataset contains metadata geolocation, object detection and tracking results. This dataset, met the expectations of our work, especially because it provides video annotations of objects tracking in different types of scenarios. The video sequences chosen are manipulated in order to improve tracking filtering quality keeping the best tracks and performing complementary annotations on video. At the end, they still contain predominantly samples of people and vehicle tracks;

Our interest is only in the position throughout time from moving object, then there is a relative decoupling of objects with the scene context. Since the dataset already offers the type object by annotations, it is unnecessary to propose an appearance model.

1) Scene modeling using mobile grid

We observed that in the fixed grid [16], the object transitions between subregions can result in a confusing global trajectory. Aiming to reduce this effect, we have different approach. Here, the subregion in the grid is also a square area with side measuring p_u pixels which is defined as grid factor. The idea is to use a mobile grid, where the leftmost region, starting each grid regions line, is positioned over the first frame data cluster by its lower right corner. From this position, the grid line is completed at the right up to the point where there are no clusters. The process is repeated for the next regions below, forming the other grid lines up to the point where no pixels or no cluster data exist in the frame. The result is a grid with a smaller numbers of areas and better positioned over the ROI, as shown in Fig. 2. The numbering of these regions is also sequential, from left to right and top to bottom. However, there is no expression to define the numbering. The motion and learning models for the mobile grid are the same as for the fixed grid. The clusters arrangement is better adjusted only in ROI where there are data samples. The clusters amount is less than fixed grid if compared with same p_u value.

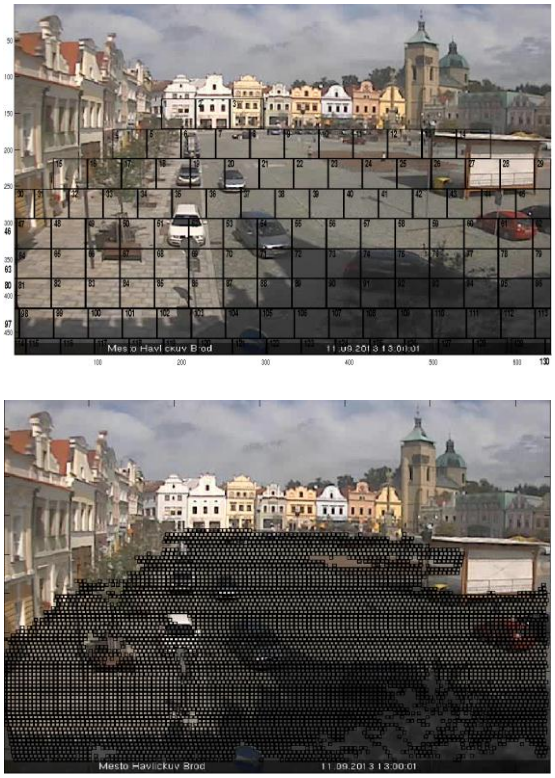

Figure 2. Mobile grid examples over a generic frame LOST video #17 when $\mathrm{p}_u = 41$ (top) and $\mathrm{p}_u = 5$ (botton).

2) Motion modeling

In the training phase, the 2D centroid coordinates of each object's trajectory are used as a reference for identify the current grid's region where is stored all data vectors generated up to a window of τ following transitions. Since the same region may belong to other paths, this region will accumulate an increasing data amount. The transitions window defines how long the object track should be observed. Here, a transition is considered when an object jump to a different region in the next observation.

The annotated dataset offers a set of u tracks T for each video, each one represented as and $k \in \mathbb{N}^*$ is the set of frames k where the object is sampled. Each frame k has a well defined *timestamp* t in the video and $t \in \mathbb{R}^+$. Then T_i^k represents a set of m observations of the same object, $T_i^k = \{O_j^k\}_j^m = 1$. Each observation is a set of transition vectors $O_j^k = \{\gamma_j^{j+a}\}_{a=1}^{\tau}$, were $\gamma_j^{j+a} = (r, v, t)^T$ is a sampled trail transition vector that contains the temporal continuous record t (*timestamp*) of object type v, in region r of the grid. Fig. 3 shows these future transitions observation of any object in frame k. They produce additional samples in the region where the object is crossing. At any track observed at any frame k, a sampling window up to τ is performed. All transition vectors up to $\gamma_j^{j+\tau}$ are associated as samples at the region in the observation point O_j^k.

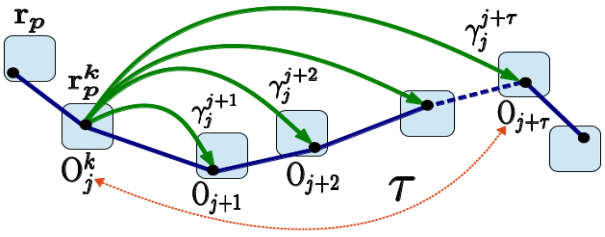

Figure 3. Detail of the motion model proposed by Basharat et al. [2] and adapted by us.

To build the database for training model, it is necessary to maintain long-term observation of the scenes in order to obtain a sufficient samples of the object types and their displacement in the scenario. For this we use two videos with different sizes, according to Table I.

TABLE I. DATA FROM LOST VIDEOS CORRESPONDING VALUES ACHIEVED AFTER TRAINING STEPS DUE TO A GRID FACTOR $p_u = 1$.

LOST dataset	video #1	video #17
resolution	480x640	480x640
hours	4	5
anormal tracks	37	116
normal tracks	1190	2990
transitions	56651	120512
samples	798048	1604119

Video #17, of greater length, has been annotated with higher normal and abnormal track amounts. The scenarios characteristics and resolutions involved were purposely chosen. The video #1 has a more sparse number of tracks than video #17. Fig. 4(a) are frame samples of video #1 and video #17 respectively. Fig. 4(b) show the corresponding normalized samples distributions reported in Table I. The pixel locations with more intense colors are those with the largest number of samples. The dispersion observed in the sample videos suggests that many areas have insufficient data for the GMM training.

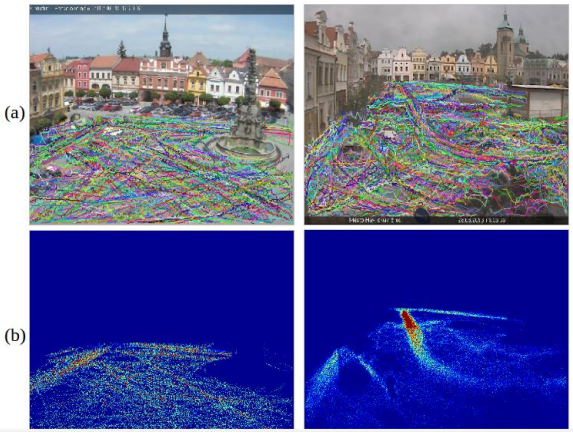

Figure 4. Detail of the scenarios used according to LOST datasets.

3) Learning modeling

Any deviation of usual local (next transition) or global (all transitions) motion results in significant differences when calculating the probability and abnormalities are so identified. Considering the data clusters dimensionality equal 3, in summary, the probability is determined through (1), where η_p represents the samples quantity in each region r_p and $a = \{1, 2, \ldots \tau\}$.

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

When $p_u = 1$ our model behaves with a pixel-based approach. Then the data in Table I shows the highest limits of sample quantities required to train our model. Since there is a mathematical relationship between the computational cost with the number of samples involved in the process, we understand it is sufficient to use the total samples as a metric to quantify and compare the results. Unlike our approach, the time complexity is much more important to motion-based approaches because of the processing of each frame subregion is continuous. In our case, once the model is trained, the decisions are computed in $O(M)$ where M depends on τ and the number of moving objects in the frame. Thus, since our goal to present a more advantageous method in terms of computational effort, we need to find the best relationship between better performance of the model as the lowest computational cost associated.

Since we are only interested in the highest hit rate of true positives (TPR) and the lowest hit rate of false positives (FPR), we adopted as reference metric the *ROC efficiency* through (2). For our binary classifier case, Powers [18] suggests a goodness performance measure for (TPR-FPR), called informedness. A number closer to 1, indicates better correct ratio for both abnormal and normal tracks. The ε value represents the number of lost tracks, which serves as penalty factor. They are represented through the numbers alongside different color

segments plotted in the ROC efficiency curves. These losses occur for two reasons: (i) the number of samples in all object transition regions was not enough for the convergence of the GMM training algorithm (usually the clusters require at least 40 samples) and (ii) lack of transitions between regions. The track loss distorts the real performance of ROC curve values because for its interpretation is considered only targets (tracks) that have at least one probability calculated from the observed transitions. In presented curves, if p_u is increased, many transition vectors are removed of the dataset, thus the track loss is increased and consequently the ROC efficiency is lower.

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

For implementation of our learning model, an iterative process is conducted in off-line mode to find the best p_u value which ensures the best performance of model. In the first step, all tracks annotated as abnormal are excluded from the dataset. The sampling for each region is performed according to the motion model. Therefore, this method is referred to as supervised learning, since the training data consists of only one class of normal events [6]. In the second step, the dataset contains normal and abnormal events so that all tracks have annotations to be used as targets to plot ROC curves. The threshold found represents the lowest probability of all transitions sampled in the scenario.

The p_u value is incremented by one from the unitary value. The limit of this increase occurs when the analysis results begin to reveal loss of original tracks or when efficiency becomes uninteresting. We observed these situations when $p_u > 30$. Fig. 5 and Fig. 7 show graphically the iterative process result, with an asterisk highlighting the best p_u values for the two evaluated videos. In the curves, the best p_u value is the best compromise among: the largest samples amount, the smallest tracks losses amount and best samples per cluster rate. The red (or darker) curve represents the samples quantity effectively used within all clusters.

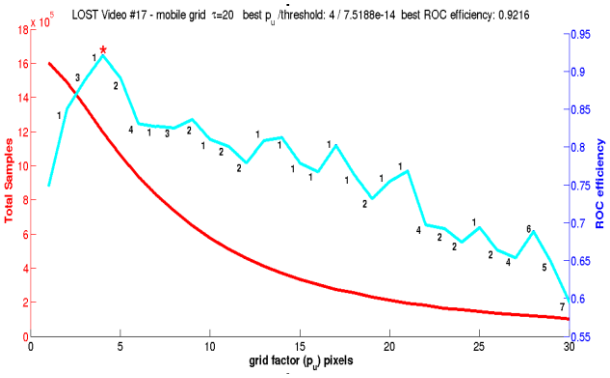

Figure 5. LOST video #17 analysis using mobile grid. Abnormal motion analysis performance (ROC Efficiency) and total samples versus p_u values variation.

As an example, Fig. 5 shows that the best ROC efficiency value occurs when $p_u = 4$. The asterisk character presents the resulting value of the (2). It also indicates the point of the best threshold value which is determined by the ROC curve. Fig. 6 illustrates the best ROC curve point also marked with an asterisk character.

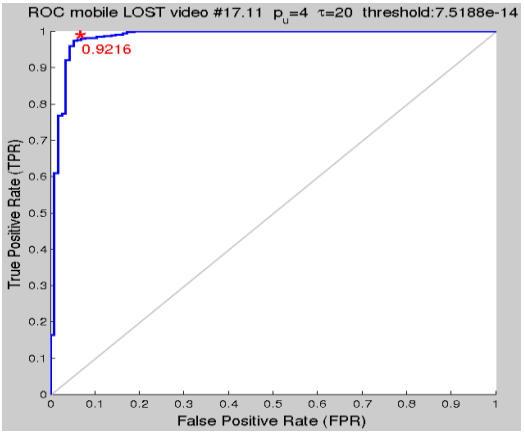

Figure 6. Best ROC curve after training process of video #17. The value under asterisk character is the ROC efficiency determined by (2).

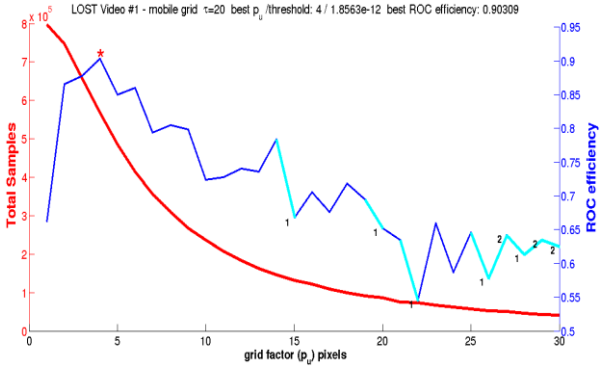

Figure 7. LOST video #1 analysis using mobile grid.

At the end of this process, we will have a searched p_u value. This value and respective best threshold ROC curve associated will be adopted for the monitored scenario. The known threshold in this off-line round will be used as a single-class classifier until the necessity of another round. Once both p_u and respective decision threshold values are chosen, any size or video sequence in the same scenario which contains the annotations on its tracking, can be tested.

As a test model, for each new position of each object in each frame, it is estimated the probability of that object type to be at the current position and time, originating from each of the τ previous transitions (high order analysis). If any of the τ probabilities is less than the threshold chosen in the learning phase, then the object is identified as describing an unusual trajectory from that point until the end of its trajectory tracking in the video. In our implementation, we highlight in red color the bounding box of the object that had its motion identified as abnormal. A screen-shot example is shown in Fig. 8.

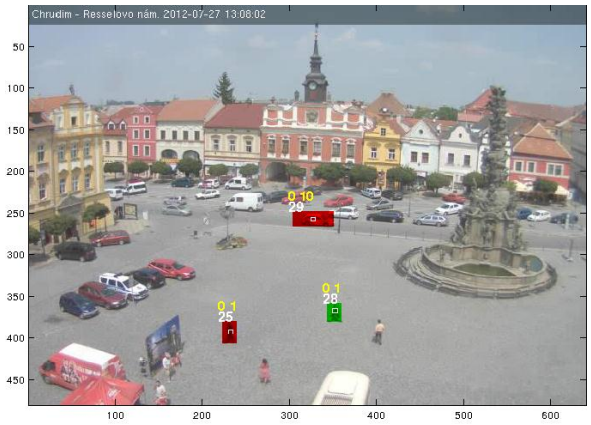

Figure 8. Screen-shot video #1 during the test phase. The bounding box with red color indicates abnormal motion. Green if normal. Track number is identified with white color. In Yellow, a owner motion and objects types labels.

In the context of this work, we have compiled the main results of the proposed method for two LOST datasets, aiming to compare with previous similar works. The proposed method was implemented in off-line mode with MATLAB, running on a computer Pentium Intel® Core™ i5 CPU M450@ 2.40GHzx4, 6GiB RAM and operating system UBUNTU12.04.

II. Results

For two datasets, the ROC efficiency is always low when $p_u = 1$ because there is a large amount of the total clusters with insufficient samples quantity for training on our learning model. This can be solved with more sample tracks. The Table II shows a summary of performance for proposals with fixed grid proposed by us [16] and mobile grid and also in relation to pixel-based model proposed by Basharat *et al.* [2]. In order to establish comparison criteria, we consider the optimal p_u grid factors for each LOST video and we use the informedness criteria (TPR-EPR) [18].

TABLE II. Main Performance Results Reached with Fixed and Mobile Grid and Equivalent Pixel-Based Approach Proposed by Basharat et al. [2].

video #1	**total samples**	**TPR-FPR**
mobile grid *best pu = 4*	567713	0.903
fixed grid *best pu = 4*	565608	0.921
pixel-based *pu = 1*	816018	0.778
video #1 7	**total samples**	**TPR-FPR**
mobile grid *best pu = 4*	1199629	0.921
fixed grid *best pu = 2*	1492117	0.878
pixel-based *pu = 1*	1646275	0.774

The performance of the mobile grid is better for the video #17 even handling a smaller amount of samples. In video #1 the mobile grid was slightly lower performance than fixed grid, due to the data are more sparse in this scenario. In addition both grid types has much better performance when compared with pixel-based model, which equates reduce $p_u = 1$. This behavior shows that the motion analysis in cluster of pixels, while reducing the computational effort, is much more effective. If we extend the comparison with the previous approach presented by Basharat *et al.* [2], the difference is huge due to the motion model these authors makes sample copies in all pixels of *bounding box* boundary. In a simulation using the dataset available by the authors, with video resolution 240x320 pixels and ~ 3 hours length, we observed more than 250 million samples and the ROC curve with much lower performance according shown in their paper.

III. Conclusions and Further Works

We present a new method for abnormal motion analysis using region-based model with mobile grid. We used datasets with video annotations aiming to isolate the motion analysis step from several pipe lines processes generally used in tracking-based approaches. The proposed region-based method supported by ROC curves and GMM, used scene, motion and learning models focused on dimensionality reduction to decrease the computational effort without sacrificing performance in detecting abnormalities. Using optimal grid size, the number of samples decreases exponentially up to ~60% if compared with equivalent pixel-based motion models, or in others words, when $p_u = 1$.

Both grid types modeling revealed similar results, however the mobile grid shows to be more accurate. The mobile grid requires more computational effort and a more elaborate algorithm. However, this additional complexity is only required once for each round of training. This slight improvement suggests that it is worthwhile to use others polygonal area forms strategies such as: Gaussianization proposed by Condurache and Mertins [19], adaptative triangular grids by Condell et al. [20] and superpixel concepts used in image segmentation area.

References

[1] S. T. Räty, "Survey on contemporary remote surveillance systems for public safety," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 5, pp. 93–515, 2010.

[2] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[3] I. Saleemi, K. Shafique, and M. Shah, "Probabilistic modeling of scene dynamics for applications in visual surveillance," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1472–1485, 2009.

[4] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011.

[5] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, and N. Tishby, "Detecting anomalies in people's trajectories using spectral graph analysis," *Computer Vision and Image Understanding*, vol. 115, no. 8, pp. 1099–1111, 2011.

[6] A. A. Sodemann, M. P. Ross, and B. J. Borghetti. "A review of anomaly detection in automated surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 42, no. 6, pp. 1257–1272, 2012.

[7] E. B. Ermis, V. Saligrama, P. M. Jodoin, and J. Konrad, "Motion segmentation and abnormal behavior detection via behavior clustering," *in Proc. ICIP*, 2008, pp. 769–772.
[8] N. Kiryati, T. Raviv, Y. Ivanchenko, and S. Rochel, "Realtime abnormal motion detection in surveillance video," in *Proc. 19th International Conference on Pattern Recognition*, 2008, pp. 1–4.
[9] Y. Shi, Y. Gao, and R. Wang, "Real-time abnormal event detection in complicated scenes," in *Proc. ICPR*, 2010, pp. 3653–3656.
[10] F. Hanapiah, A. Al-Obaidi, and C. S. Chan, "Anomalous trajectory detection using the fusion of fuzzy rule and local regression analysis," in *Proc. 10th International Conference on Information Sciences Signal Processing and their Applications*, 2010, pp. 165–168.
[11] H. Li, A. Achim, and D. Bull, "Unsupervised video anomaly detection using feature clustering," *Signal Processing, IET*, vol. 6, no. 5, pp. 521–533, 2012.
[12] A. Feizi, A. Aghagolzadeh, and H. Seyedarabi, "Behavior recognition and anomaly behavior detection using clustering," in *Proc. 2012 Sixth International Symposium on Telecommunications*, 2012, pp. 892–896,.
[13] M. Haque and M. Murshed, "Abnormal event detection in unseen scenarios," in *Proc. 2012 IEEE International Conference on Multimedia and Expo Workshops*, 2012, pp. 378–383.
[14] Y. Cong, J. Yuan, and Y. Tang, "Video anomaly search in crowded scenes via spatio-temporal motion context," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 10, pp. 1590–1599, 2013.
[15] M. Elhoseiny, A. Bakry, and A. Elgammal, "Multiclass object classification in video surveillance systems experimental study," in *Proc. CVPR'13*, 2013, pp. 788–793.
[16] J. H. B. Casagrande and M. R. Stemmer, "Region-based abnormal motion detection in video surveillance," presented at ICPRAM Angers, France, 6-8 Mar. 2014.
[17] A. Abrams, J. Tucek, J. Little, N. Jacobs, and R. Pless, "LOST: Longterm observation of scenes (with Tracks)," in *Proc. 2012 IEEE Workshop on Applications of Computer Vision*, 2012, pp. 297–304.
[18] D. M. W. Powers, "Evaluation: From precision, recall and F-factor to ROC: Informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
[19] A. P. Condurache and A. Mertins, "Accelerated nonlinear gaussianization for feature extraction," in *ICPRAM*, M. D. Marsico and A. L. N. Fred, Ed. SciTePress, 2013, pp. 121–126.
[20] J. Condell, B. Scotney, and P. Morrow, "Detection and estimation of motion using adaptive grids," in *Proc. 14th International Conference on Digital Signal Processing*, vol. 2, 2002, pp. 675–678.

Jorge Henrique B. Casagrande, graduated in Electrical Engineering (1989) and master's degree in Production Engineering (2000) from the Federal University of Santa Catarina in the concentration area of Applied Intelligence and Media and Knowledge. He is currently in the doctoral program at the Department of Automation and Systems (DAS), Federal University of Santa Catarina (UFSC). He is an effective professor (IF-SC) at the Federal Institute of Santa Catarina Campus S ão Jos é and shareholder partner of Pulso Brasil Digital Ltda. He has experience in the area of Electrical and Electronic Engineering with emphasis on Telecommunication Systems design, installation and maintenance of related equipment. Professor Casagrande has special interest in the areas of computer networks, artificial intelligence, e-business and pattern recognition applied to safety and automation systems, especially those related to computer vision.

Marcelo R. Stemmer graduated from the Federal University of Santa Catarina in Electrical Engineering in 1982, obtained his master's degree at the Graduate Program in Electrical Engineering at UFSC in the area of Control, Automation and Industrial Informatics in the period from 1983 to 1985 and his doctorate in the period 1986-1991 at the Institute WZL from the RWTH-Aachen, Germany, in the area of computer networks for industrial automation. He held a post-doctoral internship at the Institute LIP6 University Paris VI, France, in 2004. Currently serves as associate professor at the Department of Automation and Systems (DAS) at UFSC. Professor Stemmer published several articles in journals and conferences in his main areas of interest are: Industrial Networks, Computer Vision, Robotics and Artificial Intelligence.