Novel Anomalous Event Detection based on Human-Object Interactions

Rensso Mora Colque[1], Carlos Caetano[1], Victor C. de Melo[1], Guillermo Camara Chavez[2] and William Robson Schwartz[1]

[1]*Universidade Federal de Minas Gerais, DCC, Belo Horizonte, Brasil*

[2]*Universidade Federal de Ouro Preto, ICEB, Ouro Preto, Brasil*

{*rensso,carlos.caetano,victorhcmelo*}*@dcc.ufmg.br,guillermo@iceb.ufop.br,william@dcc.ufmg.br*

Keywords: Anomalous event detection, human-object interaction, contextual information.

Abstract: This study proposes a novel approach to anomalous event detection that collects information from a specific context and is flexible enough to work in different scenes (i.e., the camera does need to be at the same location or in the same scene for the learning and test stages of anomaly event detection), making our approach able to learn normal patterns (i.e., patterns that do not entail an anomaly) from one scene and be employed in another as long as it is within the same context. For instance, our approach can learn the normal behavior for a context such the office environment by *watching* a particular office, and then it can monitor the behavior in another office, without being constrained to aspects such as camera location, optical flow or trajectories, as required by the current works. Our paradigm shift anomalous event detection approach exploits human-object interactions to learn normal behavior patterns from a specific context. Such patterns are used afterwards to detect anomalous events in a different scene. The proof of concept shown in the experimental results demonstrate the viability of two strategies that exploit this novel paradigm to perform anomaly detection.

1 INTRODUCTION

Anomaly detection for video surveillance has gained importance in academy and industry. Therefore, researches have focused in extracting characteristics that help to determine anomalous events, without departing from context. This challenging task increments its hardness as semantic information of anomaly is added.

Many studies are based on a typical pipeline employing representations based on spatiotemporal features (low level characteristics extracted from temporal regions) (Hasan et al., 2016; Wang and Xu, 2016; Zhou et al., 2016; Cheng et al., 2015; Colque et al., 2015; Leyva et al., 2017) followed by one-class classification to determine whether an event is anomalous. These approaches model anomalies using characteristics such as velocity (magnitude, orientation), appearance, density and location. However, this type of information is not well-suited for solving the anomalous event detection problem since it is constrained to the same camera view, preventing the detection to be performed on different scenes within the same context. For instance, it prevents one from learning normal patterns in one particular office and detect anomalous events in another office. On the other hand, our proposed approach is based on human-object interactions for a specific context, which is more flexible and allow performing anomaly detection on different scenes within the same context.

Common representations for anomaly detection are based on the following low-level characteristics: texture (appearance), optical flow information (magnitude and orientation) and agent location. Such type of features fits well only in approaches focusing on anomaly detection for specific views, i.e., the information extracted from a particular scene cannot be used to detect anomalies in other scenes (e.g., a different camera view) because they are camera-dependent. Therefore, the investigation of approaches focused on higher-level semantic information is desired for performing a more flexible anomaly detection.

In this paper, we address the problem of anomaly detection with a different perspective. The main idea is to learn information regarding one context (e.g., office or classroom environments) from a specific scene (e.g., a particular office of classroom) and use that information to detect anomaly in other scenes belonging to the same context (e.g., a different office or classroom). In this way, our method is more flexible than the current state-of-the-art approaches found in the literature (Fang et al., 2016; Wang and Xu, 2016; Li

et al., 2016; Feng et al., 2016), which are tied to a single camera view. To be able to perform anomaly detection in multiple scenes within a given context, our proposed approach considers higher level semantic information based on human-object interactions to learn normal patterns.

Different from the current approaches that learn normal patterns based on spatiotemporal information, we learn such patterns using human-object interactions. Based on such interactions, we consider two strategies for anomaly detection: i) unrecognized interactions; and ii) incorrect sequence of interactions. While the former focuses on finding interactions that did not occur during the learning stage, the latter verifies whether the interactions occur in the same sequence as in the learning stage, otherwise they are considered as anomalies.

The main contribution of this work is a new model for anomaly representation based in human-object interactions. Our model differs significantly from the state-of-the-art approaches in how it collects the scene information. While the current models use a specific view, our model learns patterns from a scene and is able to detect anomalies in a distinct scene.

2 RELATED WORKS

A common category of the anomaly detection methods found in the literature addresses the problem by learning activity patterns from low level handcrafted visual features. In (Yu, 2014) a new feature is proposed named Mixture of Kernel Dynamic Texture (MKDT) based on (Doretto et al., 2003), a statistical model that transforms the video sequence to represent the appearance and dynamics of the video. Inspired by the classic HOF feature descriptor, a spatiotemporal feature based on both orientation and velocity was proposed in (Mora-Colque et al., 2017), which captures information from cuboids (regions with spatial and temporal support). In (Bera et al., 2016), the authors restricted their work to trajectory-level behaviors or movement features per agents, including current position, average velocity (including speed and direction), cluster flow, and the intermediate goal position. However, as mentioned in (Sabokrou et al., 2017), hand-crafted features cannot represent video events very well. Thus, our model does not use low level features as main characteristics.

Online anomaly detection is another category found in the literature. In (Javan Roshtkhari and Levine, 2013), the authors proposed an online unsupervised method, based on spatiotemporal video volume construction, using both local and global compositional information using dense sampling at various spatial and temporal scales. According to (Xu et al., 2013), the aforementioned methods model activity patterns only considering local or global context, leading to a lack of global or local information of abnormal motion pattern. In view of that, in (Xu et al., 2013) a hierarchical framework was proposed that considers both global and local spatiotemporal contexts. In (Cheng et al., 2015) was also proposed a unified framework to detect both local and global anomalies using a sparse set of STIPs. The majority of these models are based on low level features, consequently with the same advantages and drawbacks.

Recently, a growing trend is the employment of deep neural networks (DNNs). The authors in (Xu et al., 2015) proposed a novel Appearance and Motion Deep-Net (AMDN) framework for discovering anomalous activities. Instead of using handcrafted features, they learned discriminative feature representations of both appearance and motion patterns in a fully unsupervised manner. Still image patches and dynamic motion fields represented with optical flow are used as input of two separate networks, to learn appearance and motion features, respectively. After, early fusion is performed by combining pixels with their corresponding optical flow to learn a joint representation. Finally, a late fusion strategy is introduced to combine the anomaly scores predicted by multiple one-class SVM classifiers. Inspired in the same architecture, an incremental model based in Deep representation of texture and motion was proposed in (Xu et al., 2017). A spatiotemporal CNN model was developed by (Zhou et al., 2016), which accesses the appearance information and motion extracted from continuous frames. To capture anomalous events appearing in a small part of the frame.

To take advantage of the best of the two worlds (handcrafted and deep features), Hasan et al. (Hasan et al., 2016) used an autoencoder, based on the two types of features, to learn regularity in video sequences. According to the authors, the autoencoder can model the complex distribution of the regular dynamics of appearance changes. The authors in (Ribeiro et al., 2017) proposed a model that collects appearance and motion information, used to feed a Convolutional Auto-Encoder (CAE). The model is trained with normal situations uses a regularized reconstruction error (RRE) to recognize normal and abnormal events. These models focus on a specific camera view, therefore the information used for recognizing anomalies is highly correlated to the view. On the other hand, our model tries to be able to handle different views, only maintaining the same context.

The aforementioned methods perform single cam-

era view analysis, i.e., they use only one camera view to train and test. In contrast, the authors of (Loy, 2010) proposed a model based in pair wise analysis for disjoint cameras. In our study, we use contextual information to determine possible anomalous event, however using not only disjoint cameras but also different camera views and distinct environments.

3 PROPOSED APPROACH

Before presenting the proposed approach, two important aspects must be discussed: the definition used for anomalous patterns and the types of anomalies our approach intends to detect. In this work, we use the following premise: "*if something has not happened before, it will be considered as an anomalous pattern*". Hence, any pattern that differs significantly from observed patterns during the training stage should be labeled as anomalous. It is important to emphasize that such premise is very common and used on other anomaly detection methods in the literature. Regarding the types of anomaly, our model is oriented to detect anomalies in different environments, implying that typical information employed for anomaly detection (e.g., velocity and orientation captured through optical flow), cannot be used due to the change of camera position. To overcome such problem, our approach describes the scene as a set of interactions between persons and objects.

Our approach is divided in two main steps: *anomaly representation* and *anomalous pattern detection*. In the first step, a preprocessing stage allows our model to describe the sequence creating structures that represent interactions. Basically, the idea is to collect these interactions as patterns belonging to a certain person. In the second step, the extracted patterns are compared to detect whether some of them may be considered as anomalous. In the first phase, all subjects contribute by creating a set of patterns with their performed interactions. In the second phase, patterns found in training are used to detect whether someone is performing any anomaly. Thus, we can detect and locate which person in a specific frame is performing a strange or unknown activity. Here, it is important to highlight that video sequences belonging to training set are totally different from testing sequences. However, all of them have the particularity of being or having the same *context*.

3.1 Anomaly Representation

To describe the scene, our model is based on a human-object interaction representation. Therefore, we must define the scene actors and how they interact to each other. First, our model detects people and objects present in the scene using the model proposed in (Liu et al., 2015). Then, to determine which objects are interacting with a person, we employ a human pose estimation approach (Eichner et al., 2012) to locate the person's hands and to link the objects with people. Finally, we estimate people's trajectories path (tracklets) using a tracking algorithm based on Kalman filter[1].

Before describing the sequence in a set of structures, our model defines the interactions and tracklets. The first step consists on linking each person with objects. This is a challenging task, especially when depth information is not present. The proposed model uses a straightforward heuristic to link human and objects by employing the position of the hands to determine the distance between objects and the person. After each subject in the scene is linked with the interacted objects, the next step consists on tracking the person. Note that, on the first step, we can only determine unrecognized objects and, in the second step, the tracking allows our model to create a temporal representation that aims at finding unrecognized sequence interactions.

Similar to a graph, we employ a structure that represents the object interactions of a certain person in a particular frame by connecting them with edges, which we called *kernel*. Here, each interaction may be represented by a label meaning the set of objects that this specific interaction contains. Thus, we can see the tracklet representation as a list of *kernels* formed by square and circle nodes. Square nodes represent the person belonging to the tracklet, and circle nodes represent the objects that have been linked with such person. Figure 1 illustrates the representation. Note that objects may appear many times. This is done by a unique interaction label composed by the sorted object identifiers.

A set of labels that represent the *kernels* are collected. These labels are used to built a "dictionary", where each word represents a specific interaction between a person and one or more objects. For instance, interaction with the object C introduces the word "C" to the dictionary. In Figure 1, we can list the words C_1, $C_1 - C_2$, and $C_1 - C_2 - C_3$. This dictionary saves the knowledge regarding normal interactions between humans and objects that have been observed during the training stage.

[1]Since the main goal of this work is to detect anomalies exploiting a graph-based approach, we do not focus deeply on the preprocessing steps (i.e., detection and tracking).

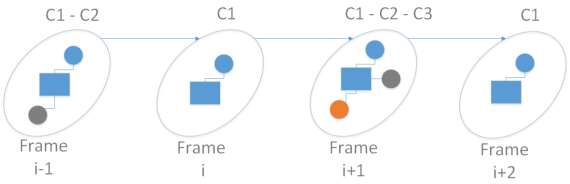


Figure 1: Interaction representation from a person tracklet. Squares represent the person at a specific frame and circles the linked objects, where different colors indicate different objects. For instance, in frame $i-1$, the blue square represents the person and the circles in blue and gray represent two different objects. The interaction with these objects creates the kernel called $C_1 - C_2$. In next frame, i, the actor interacts with a single object C_1, generating the kernel named C_1.

3.2 Anomalous Pattern Detection

Given the interaction structures, the following step is to detect anomalous patterns on the testing phase. At this stage, our model pursues two different strategies: (i) unrecognized interactions; and (ii) correct sequence interaction. The goal of these strategies is to recognize anomaly types according to the learned context information.

3.2.1 Unrecognized Interactions

In our model, this strategy represents the first level of anomaly detection, we can see it as an atomic anomaly detection. The idea is quite simple: during the training phase a list of interactions is built, then, whether an interaction is not present in this list, such node is marked as anomalous. This strategy intends to recognize when an object, or a set of objects, is present in the test but was not present in the training phase. For instance, during the training phase nobody interacted with the fire extinguisher in a computer laboratory, however, if such interaction occurs in the testing phase, this would represent an anomalous event. Figure 2 illustrates an unrecognized interaction, where the object U was not seen during the training phase, therefore it is not in the dictionary, indicating an unrecognized interaction.

3.2.2 Correct Sequence Interaction

Inspired by (Crispim et al., 2016), our second model strategy explores statistical information of sequence interactions. The idea is to detect some events according their occurrence probability, in which low probabilities indicate anomalous events. Thus, our model collects interaction information from *kernel* structures. Such information is saved in two hash tables, one of them used to count the number of occurrences for a specific word and a second used to count pairs of consecutive interactions. For instance, in Figure 1, we have two occurrences of word C_1, and one occurrence of words $C_1 - C_2$ and $C_1 - C_2 - C_3$. At same time, we have the following occurrence sequences: $\{(C_1 - C_2 \mid C_1), (C_1 \mid C_1 - C_2 - C_3), (C_1 - C_2 - C_3 \mid C_1)\}$.

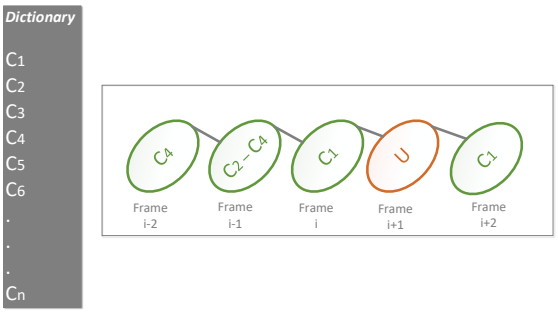


Figure 2: Example of unrecognized interaction. Frame $i+1$ illustrates the interaction with an unknown object.

Widely applied in modeling language, N-Grams utilizes the assumption that the probability of a word depends only on the previous words, this assumption is called Markov assumption. Markov models are the class of probabilistic models that assume we can predict the probability of some future unit without looking too far into the past (Jurafsky and Martin, 2009). Our model use N-grams to compute the probability of a specific sequence using maximum likelihood as following equation

$$P(w_i|w_1, w_2 ... w_{i-1}) \approx \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})} \qquad (1)$$

where, each word w_i is represented by a kernel label, i.e., the object set that interacts with the person in a specific frame. This information is found in hash tables built from interaction sequences. Finally, if the probability $P(w_i|w_{i-1})$ is less than a threshold η (0.6 in our experiments), then such interaction and consequently the observed tracklet is marked as anomalous.

Figure 3 presents a brief example of correct sequence interaction. In this case, the anomaly is given by the low probability of happening a specific sequence. In particular, the probability (Equation 1) in frame i is smaller than $\eta = 0.6$, therefore, we detect the entire tracklet as anomalous.

4 EXPERIMENTAL RESULTS

This section reports our results and is divided in two parts: (i) tests based on unrecognized interactions; and (ii) test based in anomaly sequence inter-

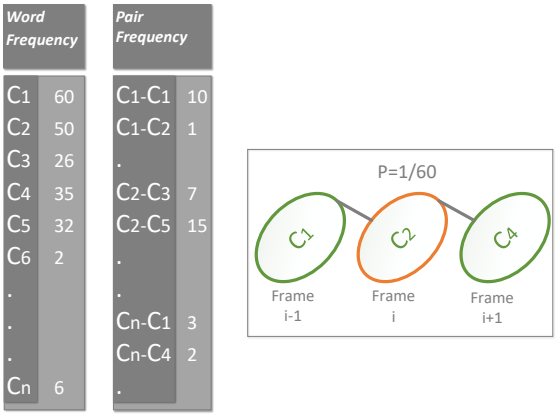


Figure 3: Example of incorrect sequence interaction. The probability in frame i is $1/60$, which would be considered anomalous for a threshold equals to $\eta = 0.6$.

actions. We use the same set of videos for training both parts.

Dataset. Our dataset was built by different views captured on different laboratories. We created a set of training videos to represent common situations, including: people interacting with chairs, laptops, backpacks and monitors. The dataset is divided in training and testing video sequences. The ground-truth was composed by every situation that was not present in the training videos. It is important to highlight the diversity not only on camera view but also in environments. Our dataset contains clips with different places and views for the same context which is a computer laboratory. For instance, various video sequences are recorded with different camera position in the room[2].

Our proposed dataset is composed of everyday laboratory work activities. For instance, students sitting and working in their computers, people entering and passing in front of the camera. Duration of the videos does not exceed five minutes of recording. It is composed by 11 clips for testing and 20 for training. As anomaly samples, we consider all interactions that a person perform with a coffee machine or a pillow (both interactions did not take place during training). We also consider anomalous sequences, the abnormal cases where a person interacted with the camera before the backpack, since in training videos people who interacted with the camera always did it with the backpack first.

Output and settings for test model. In the first stage, our model detects actors and human poses. In this case we use default CNN settings and precomputed models provided by (Liu et al., 2015) and (Eichner et al., 2012). In the second stage, we have two variables: (i) tracklet building variable $\delta_1 = 50$; and (ii) Jaccard index $j = 0.7$. The tracklet building variable is a threshold value that bounds the distance between tracklets and people occurrence. Thus, for each frame, bounding boxes are linked to a specific tracklet if the distance is smaller than this referred value (δ_1). The Jaccard index is a complement to our tracking model since in some cases the bounding boxes of certain person may vary by its pose, for instance, if a person stretches his/her arms the related bounding box will grow and only a distance threshold may give a wrong answer. In view of that, we use an occupation criteria based in overlapping areas considering the Jaccard index. To build the interaction representation for a particular tracklet, we consider a threshold $\delta_2 = 100$ to link objects with a person tracklet. Both δ_1 and δ_2 are measured in pixels.

During the test, the interaction structure is labeled as abnormal when an anomaly is detected by a of the proposed strategies. Finally, we join the anomalous events into anomaly intervals belonging to a determinate subject (tracklet). It allows us to determine which subject performed the anomaly and at what specific frame.

To evaluate the detection results, we use the metric proposed in (Cao et al., 2010). Ground truth anomalies are denoted by $Q^g = \{Q_1^g, Q_2^g, ..., Q_m^g\}$ and the output results are denoted by $Q^d = \{Q_1^d, Q_2^d, ..., Q_n^d\}$. Function $HG(Q_i^g)$ denotes whether a ground truth interval Q_i^g is detected. Function $TD(Q_j^d)$ denotes whether a detected interval Q_j^d makes sense or not. $HG(Q_i^g)$ and $TD(Q_j^d)$ are judged by checking whether Jaccard index is above a threshold th (0.30 in our experiments), according to

$$HG(Q_i^g) = \begin{cases} 1, & \text{if } \exists\, Q_k^d,\ s.t. \frac{Q_k^d \cap Q_i^g}{Q_i^g} > th \\ 0, & \text{otherwise} \end{cases}$$

$$TD(Q_j^d) = \begin{cases} 1, & \text{if } \exists\, Q_k^g,\ s.t. \frac{Q_k^g \cap Q_j^d}{Q_j^d} > th \\ 0, & \text{otherwise} \end{cases}$$

Based on the previous equations, we then compute precision and recall metrics, defined as

$$\text{Precision} = \frac{\sum_{i=1}^{m} HG(Q_i^g)}{m},$$
$$\text{Recall} = \frac{\sum_{j=1}^{n} TD(Q_j^d)}{n}.$$

4.1 Unrecognized Interactions

These experiments are oriented to determine anomalies evolving unrecognized objects. For instance, a pillow and a coffee maker usually are not in common computer laboratory and, although these objects are

[2]This novel dataset will be made publicly available after the acceptance of this paper.

not dangerous or suspicious we take in account our premise that something that not appeared during the training will be considered as abnormal.

Table 1 presents the results of such experiments. Precision and recall values are presented as tuple (P/R). We can see that the result of the tests related to the detection of anomalies due to unknown interactions presented positive results in the first level of detection. Moreover, this strategy was satisfactory to detect anomalous sequences. However, tests 3 and 4 presented a lower value of precision and recall than expected for both methods.

The reason for the low accuracy in test 3 is that we do not have a sense of depth in the image analysis. In view of that, a person with the hands far from an object in the depth can be considered as an interaction. The low recall in test 4 occurs because the pose estimator sometimes fails to detect the coordinates of the person hands. Thereby, the distance between the hands of the actor and the objects exceeds the limit to define an interaction, turning it difficult to detect the anomalies.

	Strategy 1	**Strategy 2**
Test 1	1/1	0/0
Test 2	1/0.5	1/0.5
Test 3	0.25/0	0.25/0.1
Test 4	1/0.11	1/0.07

Table 1: Precision and Recall (P/R) results for our dataset.

The idea of this experiment is to show that our model may recognize an anomalous situations given by a previously unknown interaction. We can suppose that such binary problem is easy to solve when we talk about only objects that do not appear during the training. However, it is more than only unrecognized objects since we are also looking for unknown interactions. Thus, during the training phase, interactions contain no more than three specific objects and in testing phase there are interactions with these specific objects and another that is not usually seen but appears in training phase.

For instance, interactions with laptops, chairs, notebooks and backpacks are common in computer laboratories, nonetheless, some combinations like computer and coffee cup may result abnormal, since in our premise if it not happened in training it will be anomaly in testing. Based on this, maybe there is no allowed drink in the laboratory area, however, inside the environment there is a specific coffee area. Another goal of this experiment is to determine interactions with unrecognized objects (e.g, abnormal objects, such as people interacting with weapons may be considered an abnormal situation in a laboratory).

Figure 4 shows simple examples of our model when an anomaly is detected. Subjects and objects are marked by blue and green bounding boxes, respectively. The hand positions are marked as pink circles.

4.2 Correct Sequence Interactions

Here, our main goal is to collect the temporal information about the relations in interaction structures. In this experiments the goal is to discover sequences that have not been seen before (e.g., to enter a bank office you first need to pass through a metal detector security door). In this stage, we create a synthetic situation where a video camera is placed in a particular location. Such events happened sometimes in training phase and the test case is when the device is removed. The results for this experiment are presented in Table 2. As it was expected, N-grams (Strategy 2) achieved the best results on determining anomaly sequences.

	Strategy 1	**Strategy 2**
Test 5	0/0	1/0.5
Test 6	0/0.5	0.5/0.5
Test 7	1/0.5	1/1
Test 8	0/0	1/0.33
Test 9	0/0	1/0.5
Test 10	0/0	1/1
Test 11	0/0	1/0.5

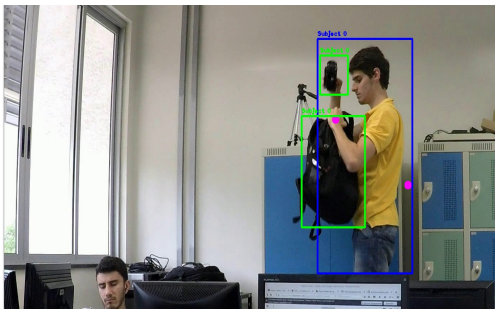
Table 2: Precision and Recall (P/R) results of our dataset.

Figure 5 shows simple examples of our model when an sequence interaction anomaly is detected. Subjects and objects are marked by blue and green bounding boxes, respectively. The hand positions are marked as pink circles. In these sequences, the context is that nobody took the camera placed in top of the locker, however during the train the student place the camera in this position.

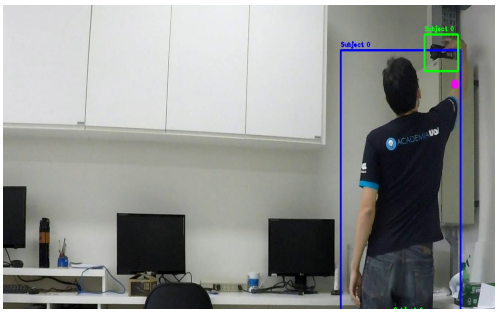
4.3 Discussion

In this section, we discuss important aspects regarding the proposed method.

Dependency on the low level tasks: It is important to highlight that our goal is to introduce a new model to recognize and detect anomalies, instead of using only a specific environment, our proposed model attempts to learn *context* based on human-object interactions. We deal with some artifacts and mistakes due to the object detector, tracking and pose recognition approaches that could be overcome by employing better approaches to generate tracklets.

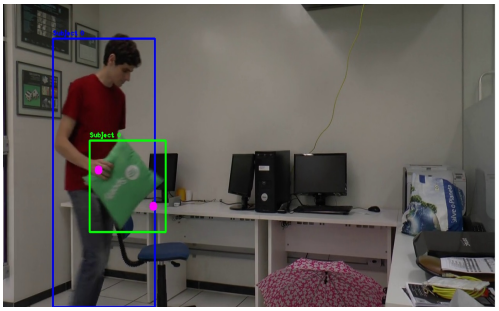


(a)

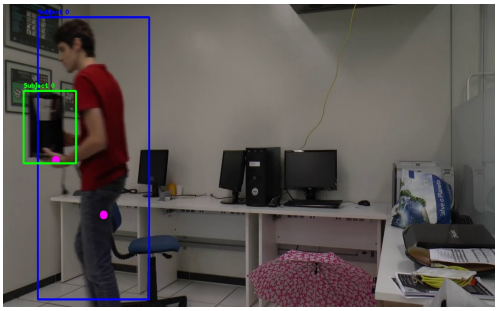


(b)

Figure 4: Sample outputs of the model. In this images unrecognized objects as: pillow and coffee maker machine was not seen in training phase. Blue bounding box correspond to a person, green for objects. Hands are marked with pink points.



(c)



(d)

Figure 5: Example of unknown sequences of interactions. The student removes camera, this action was no presented in training phase.

Human-object interaction recognition: According to experimental results, it was clear that using only euclidean distance to link humans and objects is not the best possible strategy. One possible improvement would be to use depth information to improve the link between humans and objects. However, we did not include depth information since surveillance systems generally use single camera for certain environments (turning it impossible to estimate depth).

Testing with other datasets: In literature we find several well-known datasets for anomalous event detection problem. However, such datasets are pretty specific for crowds with low resolution. Nonetheless, our model is based on human-object interactions and, in most cases of the aforementioned datasets, this type of relations are not clear given the clarity of the video sequences.

Comparisons with other approaches: As our model does not have representative results in other datasets, literature algorithms would not work properly in our proposed dataset due to the type of representations learned by them (characteristics such as speed, orientation, trajectories, appearance, textures). These types of features change significantly from scene to scene, not being representative when the environment changes.

Experimental Validation: In our experiments, we present artificial situations that could be serve as many other. Since anomalies can be seen as particular situations for a certain environment, our experiments focused on presenting the essence of our model.

5 CONCLUSIONS

In this paper, we proposed an approach for anomaly detection and localization based in context information. Instead of using common information, such as texture, magnitude or orientation, we proposed a model based in human-object interactions. An important contribution of this study is the different perspective about the information collecting and the anomaly representation. Our model was capable of detecting anomalies and to determine which individual performs an anomalous activity. As future works, we intend to focus on solving the weakness of our model presented in experiments section, such as human-object interaction in video sequences.

ACKNOWLEDGMENTS

The authors would like to thank the Brazilian National Research Council – CNPq (Grant #311053/2016-5), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project).

REFERENCES

Bera, A., Kim, S., and Manocha, D. (2016). Realtime anomaly detection using trajectory-level crowd behavior learning. In *CVPRW*.

Cao, L., Tian, Y., Liu, Z., Yao, B., Zhang, Z., and Huang, T. S. (2010). Action detection using multiple spatial-temporal interest point features. *Multimedia and Expo (ICME), 2010 IEEE International Conference on*.

Cheng, K.-W., Chen, Y.-T., and Fang, W.-H. (2015). Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In *CVPR*.

Colque, R. V. H. M., Junior, C. A. C., and Schwartz, W. R. (2015). Histograms of optical flow orientation and magnitude to detect anomalous events in videos. In *SIBGRAPI*.

Crispim, C. F., Koperski, M., Cosar, S., and Bremond, F. (2016). Semi-supervised understanding of complex activities from temporal concepts. In *AVSS*.

Doretto, G., Chiuso, A., Wu, Y. N., and Soatto, S. (2003). Dynamic textures. *International Journal of Computer Vision*.

Eichner, M., Marin-Jimenez, M., Zisserman, A., and Ferrari, V. (2012). 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision*.

Fang, Z., Fei, F., Fang, Y., Lee, C., Xiong, N., Shu, L., and Chen, S. (2016). Abnormal event detection in crowded scenes based on deep learning. *Multimedia Tools and Applications*.

Feng, Y., Yuan, Y., and Lu, X. (2016). Deep Representation for Abnormal Event Detection in Crowded Scenes. *Proceedings of the 2016 ACM on Multimedia Conference*.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., and Davis, L. S. (2016). Learning Temporal Regularity in Video Sequences. In *CVPR*.

Javan Roshtkhari, M. and Levine, M. D. (2013). An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Comput. Vis. Image Underst.*

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc.

Lab, S. V. C. (2016). UCSD anomaly data set. http://www.svcl.ucsd.edu/projects/anomaly/.

Leyva, R., Sanchez, V., and Li, C. T. (2017). Video anomaly detection with compact feature sets for online performance. *IEEE Transactions on Image Processing*.

Li, F., Yang, W., and Liao, Q. (2016). An efficient anomaly detection approach in surveillance video based on oriented GMM. In *ICASSP*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2015). SSD: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*.

Loy, C. C. (2010). Activity Understanding and Unusual Event Detection in Surveillance Videos. *Queen Mary University of London*.

Mora-Colque, R. V. H., Caetano, C., de Andrade, M. T. L., and Schwartz, W. R. (2017). Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. *IEEE Trans. Circuits Syst. Video Techn.*

Ribeiro, M., Lazzaretti, A. E., and Lopes, H. S. (2017). A study of deep convolutional auto-encoders for anomaly detection in videos. *Pattern Recognition Letters*.

Roshtkhari, M. J. and Levine, M. D. (2013). Online dominant and anomalous behavior detection in videos. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Sabokrou, M., Fayyaz, M., Fathy, M., and Klette, R. (2017). Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, pages 1992–2004.

Wang, J. and Xu, Z. (2016). Spatio-temporal texture modelling for real-time crowd anomaly detection. *Computer Vision and Image Understanding*.

Xu, D., Ricci, E., Yan, Y., Song, J., and Sebe, N. (2015). Learning deep representations of appearance and motion for anomalous event detection. In *British Machine Vision Conference (BMVC)*.

Xu, D., Ricci, E., Yan, Y., Song, J., and Sebe, N. (2017). Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Underst.*

Xu, D., Wu, X., Song, D., Li, N., and Chen, Y.-L. (2013). Hierarchical activity discovery within spatio-temporal context for video anomaly detection. In *International Conference on Image Processing (ICIP)*.

Yu, S. D. X. W. X. (2014). Crowded abnormal detection based on mixture of kernel dynamic texture. In *ICALIP*.

Yuan, Y., Fang, J., and Wang, Q. (2015). Online anomaly detection in crowd scenes via structure analysis. *IEEE Transactions on Cybernetics*.

Zhou, S., Shen, W., Zeng, D., Fang, M., Wei, Y., and Zhang, Z. (2016). Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*.