Anomalous human behavior detection using a network of RGB-D sensors

Nicola Mosca, Vito Renò and Roberto Marani, Massimiliano Nitti, Fabio Martino, Tiziana D'Orazio, and Ettore Stella

National Research Council of Italy, Institute of Intelligent Systems for Automation, Via Amendola 122/DO, 70126 Bari, Italy
mosca@ba.issia.cnr.it

Abstract. The detection of anomalous behaviors of people in indoor environments is an important topic in surveillance applications, especially when low cost solutions are necessary in contexts such as long corridors of public buildings, where standard cameras with long camera view would be either ineffective or costly to implement. This paper proposes a network of low cost RGB-D sensors with no overlapping fields-of-view, capable of identifying anomalous behaviors with respect a pre-learned normal one. A 3D trajectory analysis is carried out by comparing three different classifiers (SVM, neural networks and k-nearest neighbors). The results on real experiments prove the effectiveness of the proposed approach both in terms of performances and of real time application.

1 Introduction

Video surveillance is a rapidly growing industry. Many factors contribute to this trend such as escalating safety and security concerns, decreasing hardware costs and advances in processing and storage capabilities. In the last decade, these advances have enabled to increasingly provide automatic tools for monitoring vast areas, helping security officers in their activities[8, 11]. Traditionally, video surveillance systems have employed a network of passive ones, using fixed position and orientation cameras, sometimes assisted with pan-tilt-zoom (PTZ) enabled types. Passive camera images often require preprocessing steps designed to enable better performance, such as automatic gain and white-balance compensation, reducing issues in subsequent operations. These operations are often indispensable for addressing the challenging illumination conditions that can be found in real situations.

Video surveillance applications employ object detection algorithms, along with higher-level processing, such as tracking or event analysis, to extract meaningful data from the captured scenes. Detection algorithms vary in relation to the task to be performed and the particular context. Most of the times their focus is on moving objects in an otherwise static environment, where a background can be modelled and updated in time and moving objects are then obtained through background subtraction techniques. However, these techniques are influenced by illumination conditions, performing poorly both when images appear overly

bright and saturated, and when captured scenes are dimly lit. Artificial lights can also prove challenging since lightbulbs flicker due to alternate current, with consequences on the background modelling.

Part-based human body detectors based on color information have been proposed by [19] where SVM classifiers are used to learn specific parts of the human body on a variety of poses and backgrounds. This approach is able to handle partial occlusions enabling robust identification and tracking in crowded scenes. Nie et al. [13] developed an algorithm for tracklets association with overlapping objects for crowded scenes. Occlusions are handled using a part-based similarity algorithm while the tracklets association is formulated as a Maximum A Posterior problem by using a Markov-chain with spatiotemporal context constraints. In [5] Bouma et al. propose a system for the video surveillance of a shopping mall. In this case, the researchers employ several pedestrian detector algorithms instead of an object detector based on background subtraction, citing the limits of this technique in providing a reliable segmentation in crowded environments.

The challenges related to the use of RGB sensors, even when used with stereo algorithms, have led researchers to investigate other sensors, such as time-of-flight cameras [2], capable of directly providing depth information. The research showed the feasibility to create a people tracking system using a mean-shift algorithm for identifying interesting features aided by a Kalman filtering algorithm for predicting the next target position.

In recent times, novel camera systems such as Microsoft Kinect, pushed by research advances and economies of scale, have enabled a widespread development of 3D vision algorithms that can operate on RGB-D data [7]. Furthermore, by offloading the depth computation from the CPU to a dedicated peripheral, these systems have enabled the development of more complex techniques capable of real-time performance.

In [1] researchers proposed a multi-Kinect system designed to monitor indoor environments looking for a camera placement able to provide minimal overlapping between their field of view, in order to minimize sensor interference, a common issue in active camera systems. Positional data are expressed in a common cooordinate system enabling the whole solution to work with a combination of mean-shift and Kalman based algorithms proposed by [2] in their pipeline. Human action recognition has benefited from this trend by using techniques based on skeletal tracking. In [12] researchers used a circular array of Kinect sensors surrounding a central treadmill. Human actions are then classified by using a support vector machine operating on the extracted three dimensional skeletal data. The enhanced tracking, segmentation and pose estimation provided by Kinect libraries are used in [18] for providing accurate people segmentation. This information is then fed to a particular implementation of a Multiple Component Dissimilarity (MCD) for person re-identification through features extracted from the color data.

In addition to tracking, event analysis is another major requirement in most surveillance applications. It can be approached either with high-level semantic interpretation of video sequences or by performing anomaly detection, by

subdividing sequences in normal and a-normal sets and employing classification techniques to learn a model able to discriminate between them. In [16] Piciarelli et al. follow this approach by using a single class support vector machine able to identify anomalous trajectory.

The work presented in this paper approaches the event analysis problem by learning a model. The system, developed for the surveillance of an indoor environment, uses multiple Kinect cameras, suitably placed around a corridor for maximum coverage and no overlapping. Skeletal features are extracted from the RGB-D sensor by exploiting the OpenNi framework and by considering the extracted torso feature. A proper Kalman filter is used for the prediction step and allows robust people tracking both inter-camera and intra-camera. Trajectories are assembled together in a common reference system, by extrapolating the path using splines. Finally, anomaly behavior detection is performed using different classification algorithms by comparing multiple techniques: in addition to an SVM classifier, we use a k-nn algorithm and a feed forward neural network trained with a backpropagation algorithm.

Additional information about the methodology are reported in Section 2, while experimental results follows in Section 3. Conclusions and consideration on future researches are drawn in Section 4.

2 Methodology

The methodology proposed in this paper can be summarized in three main blocks, namely:

1. 3D Data acquisition and Preprocessing;
2. Feature Extraction;
3. Behavior Classification.

Data coming from one or multiple RGB-D sensors is initially acquired and pre-processed to obtain three dimensional trajectories of a moving subject. Then, a specific set of features is extracted from each trajectory in order to perform the classification task and understand if it leads to an anomalous behavior or not.

2.1 3D Data acquisition and Preprocessing

In the first step, several RGB-D sensors with no overlapping fields of view are employed to acquire depth data from an observed scene. In order to refer the depth maps produced by each sensor, or equivalently the corresponding point clouds, into a global reference system it is necessary to perform a preliminary calibration phase. Several reference points, whose coordinates in a global reference system are already known, are observed in each camera and are used to determine the transformation matrices between the local reference systems and the global one. In particular, knowing the position of every sample points in the local reference systems C^{Ki}, $i = 1, 2, ..n$, of the n RGB-D sensors, it is possible

to find the 4×4 matrices M_i which are able to transform every point $p^{C^{Ki}} = [x_p^{Ki}, y_p^{Ki}, z_p^{Ki}, 1]^T$, defined in C^{Ki}, into the global reference system C, since $p^C = [x_p, y_p, z_p, 1]^T = M_i \; p^{C^{Ki}}$. Solutions are obtained in the least-squares (LS) sense through the application of a standard registration algorithm, based on the single value decomposition[6].

Once all cameras refer to the same system of coordinates, people have to be detected and then tracked in time. As we will describe in the next session, in this paper we use the OpenNI framework to detect people and recover their 3D positions in each frame. Since people can move into an extended region, performing complex movements, which are not completely under the field of view of a single camera, it is necessary to put in a unique trajectory the 3D points generated by a user in each camera. For this reason a Kalman filter[3] has been designed to predict at each frame the position of the users detected at the previous frame and to further filter measurements noise.

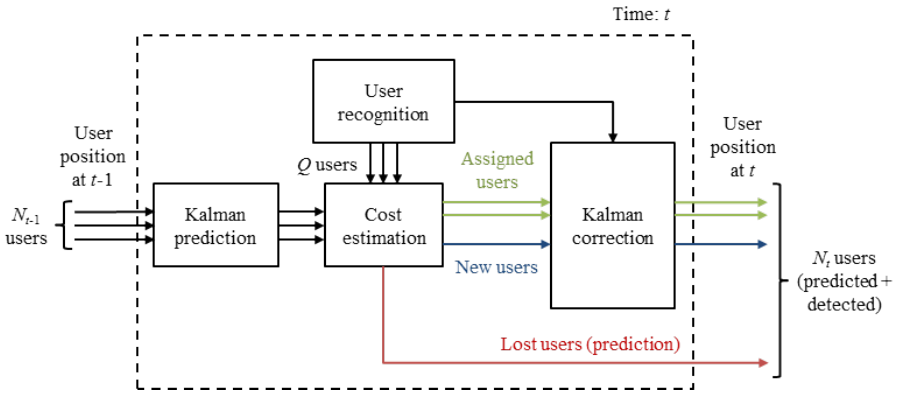

Fig. 1. Data processing scheme for user tracking with Kalman filter

Following the diagram in Fig. 1 for every frame at a specific time t, the user detection procedure segments new Q users ($Q \geq 0$) in the fields of view of the n sensors placed in the environment. On the other hand, N_{t-1} users ($N_{t-1} \geq 0$) were computed at the previous discrete time instant $t-1$. The task of user tracking aims to associate, if possible, users detected at time t with those identified at time $t-1$.

We suppose that each user detected at $t-1$ moves with constant velocity. Its position is thus predicted by using a Kalman filter, which operates over a state vector defined by the position and the speed of the user. The predicted positions of the N_{t-1} users are thus compared with those observed in the environment. This comparison is mediated by a cost computation, easily defined in terms of the Euclidean distance between the positions of every current user and the N_{t-1} previous ones. Users with close positions, i.e. with small cost values, are in rela-

tion. Finally, for each reassigned user, the state of the Kalman filter is updated in order to reduce the contribution of measurement noise. This strategy is applied between every pair of consecutive frames, where in general three different events can arise:

- Users are still visible in the field of view of the specific sensor and thus are correctly assigned to corresponding new instances observed in the scene. In this case the Kalman filter operates to correct measurement, in accordance with the previous estimation;
- New users enter in the scene and are detected in the current frame t. New instances are then initialized with the states of the detected users;
- Users are lost and no longer visible in the fields of view of the sensors. The states of the lost users are still kept in the analysis and evolve following the model of the Kalman filter, i.e. at constant velocity.

As a result of this processing, each user is tracked within the whole environment leading to the generation of a trajectory $\Theta_j = [\theta_1, \theta_2, \ldots, \theta_{N_j}]$ that contains N_j 3D points. Hence, θ_k represents the 3D information associated at time t_k. The number of points N_j depends on the duration of the time interval in which the specific user U_j is tracked by the proposed algorithm.

Moreover, each trajectory has been fit on a smoothing spline σ to obtain a single continuous trajectory starting from multiple sub-trajectories acquired from each sensor. σ is a curve defined starting from a smoothing parameter s (in our experiments $s = 0.99$) so that the following quantity gets minimized:

$$s \sum_i (\Theta(t_k) - \sigma(t_k))^2 + (1 - s) \int (\frac{d^2\sigma}{dt^2})^2 dt$$

where t_k represents the time in which a point is observed or interpolated. Both $\Theta(\cdot)$ and $\sigma(\cdot)$ are referred to the same time basis.

2.2 Feature Extraction

Trajectories can be seen as raw data that need to be managed by the classifier to understand whether a behavior of the selected user is anomalous. This goal can be achieved by creating a more discriminative representation of the trajectories, i.e. feature vectors. In this case, eleven features have been identified for each trajectory and have been used to define the feature vector $x = [x_1, x_2, \ldots, x_{11}]$ that will be the input of the subsequent classifier. x is populated in the following manner:

- the first five elements are, respectively: mean, median, standard deviation, median absolute deviation (MAD) and maximum value of the velocity computed on Θ_i (defined as the ratio of the difference of position on the XY plane and the temporal difference between subsequent frames);
- the next five elements are: mean, median, standard deviation, MAD and maximum value of curvatures that have been evaluated on the spline trajectory σ_i. Each curvature is defined as the reciprocal of the radius of the circumference that passes through three consecutive trajectory points;

– the last element of the feature vector is the number of trajectory intersections with itself.

2.3 Behavior Classification

Three different supervised classifiers have been employed in this experiment: a support vector machine (SVM)[4], a k-nearest neighbor (k-NN)[9] and a neural network (NN)[10]. The first one is a binary classifier that tries to estimate the boundary that best divides two different clusters in a multidimensional space. In other words, it looks for the hyperplane that minimizes the distance with respect to training data by solving an optimization problem. K-nearest neighbor classifies a new incoming sample evaluating its k nearest samples among training data by means of a voting procedure (in this case, k has been set to 1). Finally, the neural network tries to approximate a model of an unknown function by using artificial neurons arranged in several layers and changing the weights of the connections between them. In this work, 11 input features are mapped on two classes (normal, not normal) mapped on two different nodes. This way ambiguity cases can be detected and dealt accordingly.

3 Experiments and discussions

The next sub-sections will introduce the actual setup used in our experiments, describing the implemented sensors and the system architecture. Input dataset will be presented together with classification results obtained by SVM and K-NN and Neural Network.

3.1 Experimental setup

The proposed methodology has been applied to the analysis of videos produced by a set of RGB-D camera placed within an indoor environment, namely a corridor. With reference to the sketch map in Fig. 2a, three Microsoft Kinect sensors K_1, K_2 and K_3 are arranged within the corridor. Specifically, K_2 and K_3 focus on the boundaries of the corridor, while K_1 looks at the central area (Fig. 2b). Each sensor is locally connected to a node for the data storage, whereas the whole system is remotely controlled by a server unit via a UDP protocol. The server sends a start signal to every node which enable video recording. Each video, which lasts 30 seconds, is finally downloaded by the server. A start signal is sent to the nodes through the network and thus is received with slight delays. However this is negligible and does not affect the whole system pipeline.

As an example, a frame captured by the Kinect K_2 is displayed in Figs. 2c-d, where a depth map and the corresponding RGB image are shown, respectively.

The position of the three Kinect cameras has been set in order to cover the highest area without overlapping of their cones of sight (red regions in the sketch). It ensures the best working conditions for the sensors, since no interference phenomena would alter the depth maps. However, it produces shadow

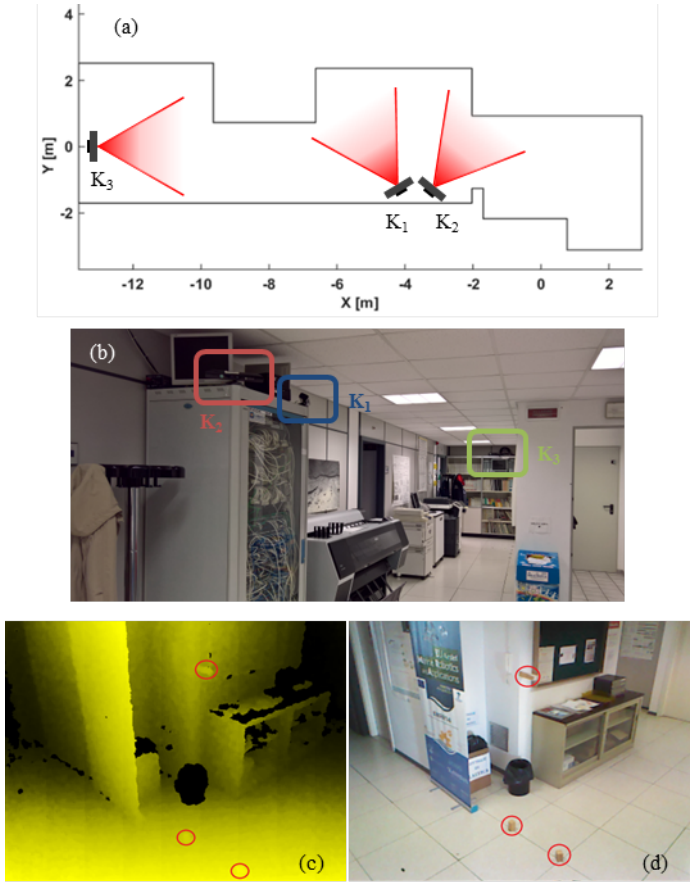

Fig. 2. (a) Map of the corridor and position of the three Microsoft Kinect cameras used in the proposed experiments. (b) Picture of the actual environment. (c)-(d) Depth map and corresponding RGB image. Red circles highlight objects in actual relationship.

areas, e.g. the regions between K_1 and K_2 (narrow shadow) or between K_1 and K_3 (wide shadow). As previously stated, the calibration phase needs a few points of known position in the reference systems of both the kinect cameras and an external surveying instrument.

In figure 2c-d the red circles enclose corresponding objects between the depth and the RGB images, which are used to calibrate sensors and transform data into a global reference system. In order to measure the position and attitude of each camera a theodolite (Nikon Total Station D50[14]) has been used.

In this paper, in order to detect and segment the people silhouettes we have used the well known OpenNI framework together with the Primesense NiTE library[15], which is able to recognize and track up to 15 user skeletons. Although, this framework also integrates a robust algorithm for tracking, it has been used

only for the extraction of the skeletal joints, specifically the torso joint, which is assumed as the center of mass of the detected user. Additionally, users can perform complex movements subtending different cameras. Since each Kinect works independently from the others, it is necessary to address people re-identification on a higher level, which is not provided by OpenNI.

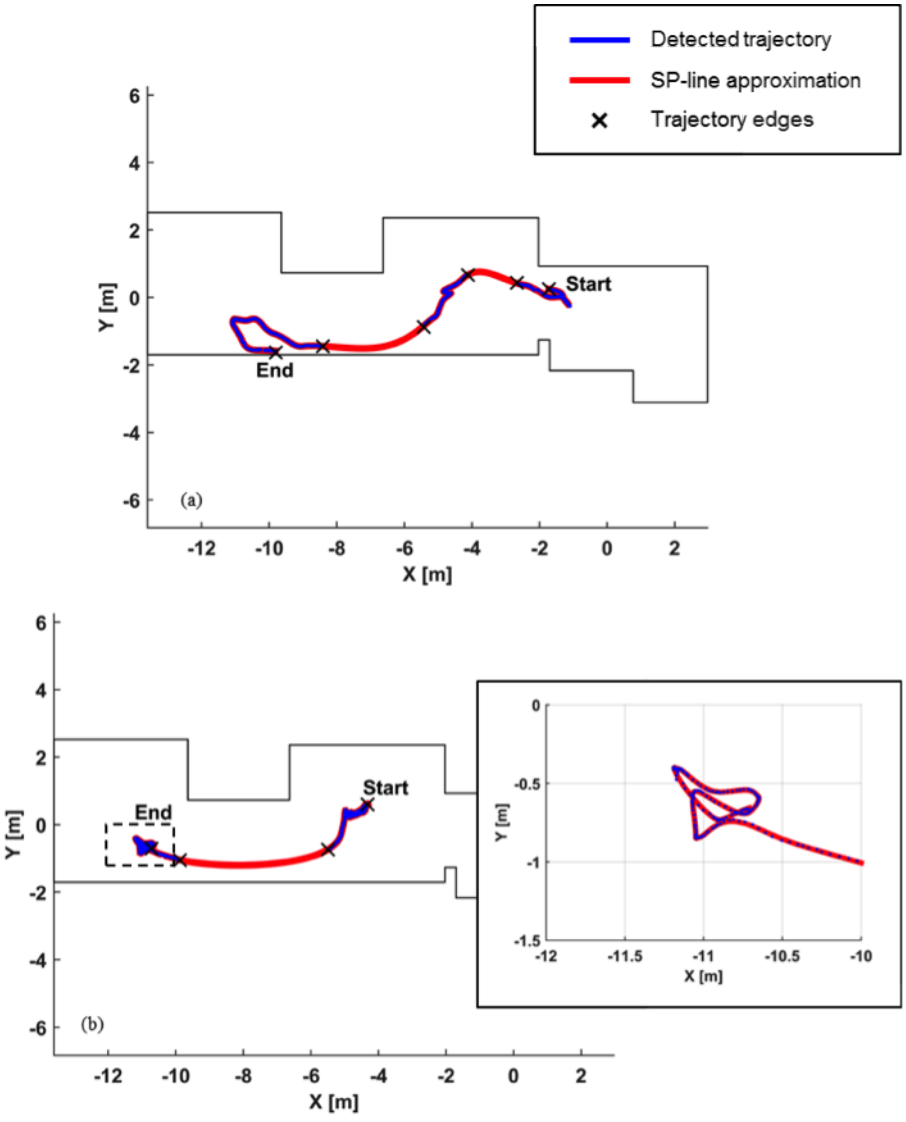

Fig. 3. Comparison of trajectories belonging to (a) normal and (b) anomalous behavior classes. The inset of (b) highlights the final part of an anomalous trajectory.

Two examples of acquired trajectories belonging to the two different classes of behavior (normal and anomalous) are reported in Fig. 3, with a single user moving with-in the environment. Here, blue lines display the actual trajectories captured by the RGB-D cameras, whereas the red ones are those generated by spline interpolation, which is also able to reconstruct the user movements out of the fields of view of the three Kinect sensors.

Table 1. Confusion matrices and average accuracy value for the experiment. Each entry of the table represents a confusion matrix in which diagonal elements represent the correct prediction while off-diagonal elements are classification errors. Accuracy per each run is reported in green, and the last column shows the average value of the accuracy achieved by each classifier in the three experiments. The best one is highlighted in bold and corresponds to the neural network one: 93.9%.

	Run 1			Run 2			Run 3			Accuracy
SVM		0	1		0	1		0	1	90.5%
	0	53.3%	1.7%	0	53.3%	1.7%	0	53.3%	1.7%	
	1	6.7%	38.3%	1	8.3%	36.7%	1	8.3%	36.7%	
	91.6%			90%			90%			
K-NN		0	1		0	1		0	1	87.7%
	0	50%	5.0%	0	48.3%	6.7%	0	50%	5.0%	
	1	6.7%	38.3%	1	6.7%	38.3%	1	6.7%	38.3%	
	88.3%			86.6%			88.3%			
NN		0	1		0	1		0	1	**93.9%**
	0	53.3%	1.7%	0	55.0%	0%	0	55.0%	0%	
	1	1.7%	43.3%	1	6.7%	38.3%	1	8.3%	36.7%	
	96.6%			93.3%			91.7%			

3.2 Classification results

The preliminary task of trajectory extraction has been used to create a dataset of 60 user paths within the corridor under inspection. Each path refers to the observation of a single individual that has been recognized among the three Kinect sensors. It should be noted that user paths are extracted also when many people moves simultaneusly in the scene, as the tracking procedure based on the Kalman filter prediction is able to disambiguate the great majority of people intersection.

In the whole dataset, 33 trajectories are associated with a normal behavior and are labeled with 0 (the 55%), while the remaining 27 anomalies are associated

to the value 1 (the 45%). A k-fold cross validation method (with $k = 5$) has been employed to evaluate the capabilities of all the classifiers on the entire available data, since training and test set change and span the whole dataset. For this reason, data has been randomly partitioned in 5 subsets to build the training set with 80% of data and the test set with the remaining 20%. Then, training and testing tasks are repeated 5 times per run, iteratively changing the test set with one of the partitioned subsets. Moreover, in order to better evaluate the accuracy of the tested classifiers, the experiment has been repeated three times per each classifier by changing the initial condition (random seed) used for partitioning.

Results are reported in Table 1. Three experiments are repeated for each classifier, for a total of 9 confusion matrices. The last column reports the average accuracy value for the three runs. The accuracy of each run has been shown in green, under the confusion matrices. The first thing to notice is that both SVM and neural network are able to exceed 90% accuracy value, implying that the chosen features show acceptable discriminating capabilities when used with such classifiers. On the contrary, K-NN has the worst performances among the classifiers. In particular, the neural network seems to be the best classifier among those considered, as it produces results on average around the 93.9%.

Some examples of the classified paths are reported in figure 4. On the left three normal behaviors correctly classified, while on the right three anomalous behaviors, which are characterized by repeated changes of the directions or long periods of standing still.

4 Conclusions

In this paper we propose the use of multiple Kinect cameras for developing a low cost surveillance system able to recognize anomalous human behavior. The torso node is extracted from the skeleton features provided by the OpenNi framework. A proper Kalman filter is used for the prediction step and allows robust people tracking both inter-camera and intra-camera. Anomalous behavior detection is performed using different classification algorithms by comparing multiple techniques: ANN, SVM and k-nn.

Experimental results demonstrate that the proposed architecture and the developed methodologies are able to recognize anomalous behavior in the majority of cases with respect to the total of observed path. However, it should be noted that the initial association of the paths in the dataset to normal and anomalous behaviors has been done by a human operator observing each path performed by the users. In future researches, more paths will be considered simultaneously to make a decision about a behavior, as those considered as anomalies could be only due to interactions among people.

In its actual form, the system could fail when multiple people enter simultaneously in each camera field of view and do not maintain the same walking direction when intersects or cross the occluded areas of the scene among different cameras. In this case a people re-identification procedure [17], based on

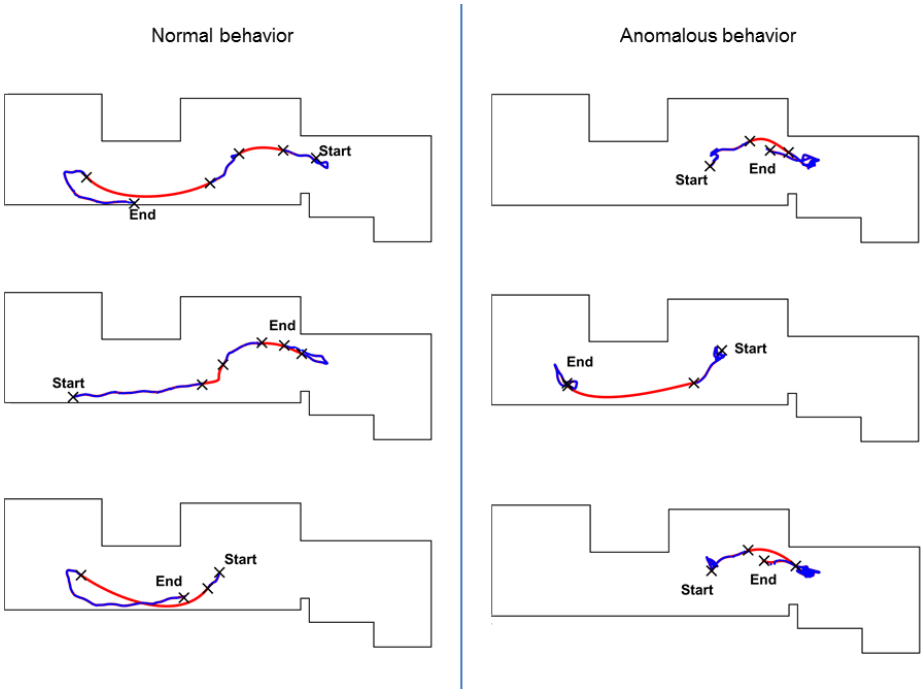

Fig. 4. Examples of classified trajectories: on the left normal behavior, on the right anomalous behavior.

color features could be used to avoid false associations and perform correctly the trajectory reconstruction.

References

1. Almazan, E., Jones, G.: Tracking people across multiple non-overlapping rgb-d sensors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 831–837 (2013)
2. Bevilacqua, A., Di Stefano, L., Azzari, P.: People tracking using a time-of-flight depth sensor. In: 2006 IEEE International Conference on Video and Signal Based Surveillance. pp. 89–89. IEEE (2006)
3. Bishop, C.M.: Pattern recognition. Machine Learning 128 (2006)
4. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 144–152. ACM (1992)
5. Bouma, H., Baan, J., Landsmeer, S., Kruszynski, C., van Antwerpen, G., Dijk, J.: Real-time tracking and fast retrieval of persons in multiple surveillance cameras of a shopping mall. In: SPIE Defense, Security, and Sensing. pp. 87560A–87560A. International Society for Optics and Photonics (2013)
6. Chen, Y., Medioni, G.: Object modelling by registration of multiple range images. Image and vision computing 10(3), 145–155 (1992)

7. D'Orazio, T., Marani, R., Renò, V., Cicirelli, G.: Recent trends in gesture recognition: how depth data has improved classical approaches. Image and Vision Computing (2016)
8. D'Orazio, T., Guaragnella, C.: A survey of automatic event detection in multi-camera third generation surveillance systems. International Journal of Pattern Recognition and Artificial Intelligence 29(01), 1555001 (2015)
9. Fix, E., Hodges Jr, J.L.: Discriminatory analysis-nonparametric discrimination: consistency properties. Tech. rep., DTIC Document (1951)
10. Haykin, S., Network, N.: A comprehensive foundation. Neural Networks 2(2004) (2004)
11. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 34(3), 334–352 (2004)
12. Kwon, B., Kim, D., Kim, J., Lee, I., Kim, J., Oh, H., Kim, H., Lee, S.: Implementation of Human Action Recognition System Using Multiple Kinect Sensors, pp. 334–343. Springer International Publishing, Cham (2015)
13. Nie, W., Liu, A., Su, Y.: Multiple person tracking by spatiotemporal tracklet association. In: Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on. pp. 481–486. IEEE (2012)
14. Nikon: Total station, http://www.nikon.com/about/technology/life/others/surveying/
15. OpenNI: Openni website, http://openni.ru/
16. Piciarelli, C., Micheloni, C., Foresti, G.L.: Trajectory-based anomalous event detection. IEEE Transactions on Circuits and Systems for video Technology 18(11), 1544–1554 (2008)
17. Renò, V., Politi, T., D'Orazio, T., Cardellicchio, A.: An human perceptive model for person re-identification. In: VISAPP 2015. pp. 638–643. SCITEPRESS (2015)
18. Satta, R., Pala, F., Fumera, G., Roli, F.: Real-time appearance-based person re-identification over multiple kinecttm cameras. In: VISAPP (2). pp. 407–410 (2013)
19. Shu, G., Dehghan, A., Oreifej, O., Hand, E., Shah, M.: Part-based multiple-person tracking with partial occlusion handling. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 1815–1821. IEEE (2012)