technologies

MDPI

Article

Fall Detection Using Multi-Property Spatiotemporal Autoencoders in Maritime Environments

Iason Katsamenis *, Nikolaos Bakalos, Eleni Eirini Karolou, Anastasios Doulamis and Nikolaos Doulamis

School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, 15780 Athens, Greece; bakalosnik@mail.ntua.gr (N.B.); ele.karolou@gmail.com (E.E.K.); adoulam@cs.ntua.gr (A.D.); ndoulam@cs.ntua.gr (N.D.)

* Correspondence: iasonkatsamenis@mail.ntua.gr

Abstract: Man overboard is an emergency in which fast and efficient detection of the critical event is the key factor for the recovery of the victim. Its severity urges the utilization of intelligent video surveillance systems that monitor the ship's perimeter in real time and trigger the relative alarms that initiate the rescue mission. In terms of deep learning analysis, since man overboard incidents occur rarely, they present a severe class imbalance problem, and thus, supervised classification methods are not suitable. To tackle this obstacle, we follow an alternative philosophy and present a novel deep learning framework that formulates man overboard identification as an anomaly detection task. The proposed system, in the absence of training data, utilizes a multi-property spatiotemporal convolutional autoencoder that is trained only on the normal situation. We explore the use of RGB video sequences to extract specific properties of the scene, such as gradient and saliency, and utilize the autoencoders to detect anomalies. To the best of our knowledge, this is the first time that man overboard detection is made in a fully unsupervised manner while jointly learning the spatiotemporal features from RGB video streams. The algorithm achieved 97.30% accuracy and a 96.01% *F*1-score, surpassing the other state-of-the-art approaches significantly.

Keywords: man overboard; deep learning; computer vision; unsupervised learning; convolutional autoencoder; spatiotemporal data

Citation: Katsamenis, I.; Bakalos, N.; Karolou, E.E.; Doulamis, A.; Doulamis, N. Fall Detection Using Multi-Property Spatiotemporal Autoencoders in Maritime Environments. *Technologies* **2022**, *10*, 47. https://doi.org/10.3390/technologies10020047

Academic Editor: Fillia Makedon

Received: 15 January 2022
Accepted: 25 March 2022
Published: 29 March 2022

1. Introduction

A man overboard is an emergency incident where a passenger or a member of the ship's crew has fallen off the ship into the sea and requires immediate rescue. It is underlined that, every year, about 22 people on average fall off a cruise ship, and 79% of them do not survive or are considered missing [1]. The principal cause for such a small survival rate is the fact that, when a person remains for one hour in water at 4.4 °C, their body temperature drops to 30 °C [2]. Thereby, man overboard is a serious and critical event that requires immediate handling, since the overboard casualty is exposed to numerous safety hazards, such as drowning, injuries, hypothermia, and rough sea. This entails that time is a crucial factor and plays an important role in the effective recovery of the victim, and therefore, the lack of timely and critical information regarding the man overboard occurrence (e.g., exact time and location of the event) can lead to unpleasant and serious consequences [3].

Conventional maritime surveillance systems contain optical grayscale or RGB sensors that are programmed to monitor various predetermined locations of the safety perimeter [4]. It should be noted, however, that such traditional surveillance methods require an operator who has to monitor simultaneously several real-time video sequences. This can undeniably lead to an increased chance of error and, thus, inadequate response to the critical event. Therefore, for the effective evaluation and localization of the critical event, a continuous real-time monitoring approach should be adopted.

Recent developments in signal processing and deep learning applications of imagery from various optical sensors have received much attention from the scientific community, as well as brought about significant progress and remarkable breakthroughs in the field of intelligent video surveillance [5,6]. This fact has led algorithms to be a powerful tool in various monitoring systems that are based on semantic information extraction (e.g., anomaly detection [7], motion tracking [8], pose estimation [9,10], and human detection [11,12]). Moreover, the recent abundance of large datasets and high-quality RGB cameras, as well as tremendous developments in computer power and cloud computing, have led deep neural networks to play a crucial role in intelligent surveillance systems [3,13–16].

The present paper outlines a deep learning approach of effectively recognizing man overboard falls in video sequences captured from multiple RGB cameras that are installed at the ship's perimeter. It is underlined that, due to the lack of man overboard data and, at the same time, the abundance of datasets that are related to the normal situation and activities, we address the fall identification task as an anomaly detection problem. Thereby, the aim of our research is to broaden the current knowledge of unsupervised learning and anomaly detection and train a multi-property spatiotemporal convolutional autoencoder framework based on the normal conditions, as well as utilize their reconstruction error to recognize man overboard as an abnormal event, during the validation procedure.

It is noted that fully connected, as well as deep autoencoders, ignore 2D image structures, a fact that introduces redundancy in the parameters, forcing each feature to be global and thereby span the entire visual field [17]. On the contrary, convolutional autoencoders outperform the aforementioned conventional techniques in various computer vision tasks, since they are able to discover localized features that repeat themselves all over the input [17]. In parallel, since a video sequence incorporates information in both time and space, a spatiotemporal convolutional autoencoder is more suitable for the man overboard recognition problem. This is mainly due to the fact that it is able to learn a representation of the local spatiotemporal patterns of frames in a video stream [18,19].

Nevertheless, modern automated approaches, as described in the previous paragraphs, present a series of crucial drawbacks. Firstly, supervised learning approaches fail to generalize; thus, their performances tend to be reduced significantly when the general setting of the application (e.g., the background of the falling event) changes. Moreover, a supervised approach requires large, annotated datasets; hence, there are performance bottlenecks related to the number of annotated samples, the captured scenarios, etc. Regarding the unsupervised techniques, the approaches that present the best performance utilize action recognition preprocessing steps in order to process the action and provide the analysis. This results in the need for multiple forward propagations in learning structures. It is underlined that, while such approaches are essential in a setting with numerous humans present who are moving in multiple trajectories, they do not capture the particularities of the studied application scenario (i.e., man overboard identification).

To this end, in this study, we present a novel man overboard detection framework that includes: (i) formulating the identification of the critical event as an anomaly detection task; (ii) designing a set of spatiotemporal convolutional autoencoders over multiple image properties (i.e., appearance, gradient, and saliency) of the RGB data; and (iii) training the autoencoders on the normal situation and utilizing their reconstruction error to detect unseen man overboard events during testing. It is highlighted that the proposed method manages to identify falls by using only one forward propagation on our learning architecture. This paper is an extension of our previous study [15], and to the best of our knowledge, this is the first time a deep spatiotemporal convolutional autoencoder has been utilized for the man overboard identification problem from RGB video sequences.

The remainder of this paper is organized as follows. Section 2 briefly presents human detection frameworks integrated into intelligent video surveillance systems. Section 3 describes the overall architecture of the proposed methodology. Section 4 analyzes the experimental results obtained. Section 5 briefly discusses various aspects and potential

improvements of the proposed deep learning framework. Lastly, in Section 6, the conclusion is summarized, and suggestions for future research are introduced.

2. Related Work

Recently, with the advancement of deep neural networks, intelligent video surveillance systems have gained significantly increased interest in the computer vision community [11]. The key feature of a universal maritime surveillance application is human recognition, and thus, it must be completely independent of the environment, as well as weather and light conditions [3]. Several methods for detecting humans through optical imaging have been presented in the literature. More specifically, many of them utilize features extracted from the histograms of oriented gradients [20], in conjunction with various classification techniques (e.g., support vector machines [21], AdaBoost [22], and *k*-means [23]). In parallel, several studies have proposed various techniques for the human detection problem that exploit the probabilistic assembly of robust part detectors [24], depth information [25], classification on Riemannian manifolds [26], and flexible mixture of parts [27].

More recent evidence on this topic highlights that the rapid development of deep learning models has brought significant progress into the field of intelligent video surveillance from RGB data [28,29]. It is underlined that multiple studies have outlined the significance of real-time home surveillance applications [30,31], which focus on fall detection through RGB optical sensors, computer vision, and deep learning frameworks [32–34]. Table 1 provides a tabulated summary of deep learning techniques that utilize RGB data developed for a more general fall detection scenario. Nevertheless, little work has been presented in the literature on maritime surveillance systems and, in particular, on the man overboard incident [3].

Table 1. Summary of the state-of-the-art machine learning techniques employed for human fall detection.

Authors	Utilized Deep Learning Technique	Utilized RGB Dataset
Abobakr et al. [35]	CNN, RNN, LSTM with ResNet, Recurrent LSTM, and Logistic regression	URFD dataset
Adhikari et al. [36]	CNN	Own dataset
Cameiro et al. [37]	CNN	URFD and FDD dataset
Espinosa et al. [38]	CNN	UP-Fall and Multicam dataset
Ge et al. [39]	RCN, RNN, and LSTM	ACT4^2 dataset
Hsieh and Jeng [40]	FOF CNN and 3D-CNN	KTH dataset
Hwang et al. [41]	3D-CNN	TST Fall detection dataset
Kasturi et al. [42]	3D-CNN	URFD dataset
Li et al. [43]	CNN	URFD dataset
Li et al. [44]	3D-CNN	Own dataset
Lie et al. [45]	CNN, RNN, LSTM, and DeeperCut	Own dataset
Lin et al. [46]	RNN and LSTM	Own dataset
Lu et al. [47]	CNN	URFD, FDD, and Multicam dataset
Lu et al. [48]	3D-CNN and LSTM	Sports-1M and Multicam dataset
Rahnemoonfar and Alkittawi [49]	3D-CNN	SDUFall dataset
Shen et al. [50]	DeepCut	Own dataset
Tao and Yun [51]	RNN and LSTM	Rougier and Meunier dataset [52]
Tsai and Hsu [53]	CNN (MyNet1D-D)	NTU RGB+D dataset
Zhou and Komuro [54]	Variational Auto-encoder	HQFD and Le2i dataset
Zhou et al. [55]	CNNs based on AlexNet and SSD-Net	Own dataset

As already mentioned, the man overboard event can be addressed as a computer vision problem of detecting abnormal behavior. In particular, the normal situation comprises the typical capture of the ship's surroundings, while the anomaly could be the capture of a human fall. Various unsupervised learning techniques have been proposed for abnormal event detection. The work of [56] introduced a framework for anomaly detection that is independent of the temporal ordering of anomalies and unsupervised, thus requiring no separate training sequences. Other studies exploit the online detection of abnormal events using incremental coding length [57] and unmasking, a technique previously used for authorship verification in text documents [58]. In parallel, the works of [59,60] employed tracking algorithms in order to extract salient motion information that is subsequently classified as normal or abnormal. It was underlined, however, that in complex visual environments (e.g., scenes where numerous humans are present), the tracking procedure tends to fail.

Furthermore, an increasing number of studies have utilized deep convolutional autoencoders to detect anomalies on videos [61,62]. Other works utilize convolutional autoencoders in combination with a one-class support vector machine for video anomaly detection [63,64]. The authors of [65] leverage the conventional handcrafted spatiotemporal local features and learn a fully connected autoencoder on them and then build a fully convolutional feed-forward autoencoder to learn both the local features and the classifiers as an end-to-end learning framework. The work of [66] proposes a technique to generate unbiased features by unsupervised learning for detecting irregularities in high-dimensional data feed (e.g., surveillance system for industrial robots). In [67], the authors introduced a novel double-fusion framework, exploiting the complementary information of both appearance and motion patterns, to detect anomalies in video sequences.

Several studies have been carried out demonstrating spatiotemporal autoencoders, which utilize 3D convolutions to extract spatiotemporal features from video sequences, in order to find anomalies in various types of data (e.g., RGB [68], thermal [16], and hyperspectral [69]). The authors of [70] proposed a two-stage cascade of classifiers for anomaly detection and localization in video data, where spatiotemporal patches are fed into a 3D autoencoder for the initial identification of regions of interest and then evaluated in the second stage by a more complex and deeper 3D convolutional neural network. In parallel, the work of [71] proposed a hybrid framework, based on the Long Short-Term Memory (LSTM) encoder–decoder and the convolutional autoencoder, which not only extracts better spatiotemporal context but also improves the extrapolate capability of the corresponding decoder with the shortcut connection. Lastly, deep generative models have been proposed in several works for detecting abnormal events in various complex environments [65,71,72].

Our Contribution

Inspired by the above research work, in the present study, an unsupervised fall detection technique for identifying man overboard events is utilized. The proposed framework (see Figure 1) is based on a spatiotemporal convolutional autoencoder, which is trained on RGB video sequences that simulate man overboard scenarios. Our network is trained on the normal situation in order to learn efficient data encodings by ignoring signal noise and then uses its reconstruction error to detect man overboard as an abnormal event during the test process. In parallel, we utilize multiple image properties (i.e., appearance, gradient, and saliency) of the RGB data to enhance the identification capabilities of the proposed architecture. To the best of our knowledge, man overboard identification has not been addressed as an anomaly detection task utilizing unsupervised deep learning techniques. Lastly, in this study, we present a dataset containing RGB videos with test throws of a human-sized dummy that was created to train and validate the performance of the proposed framework.

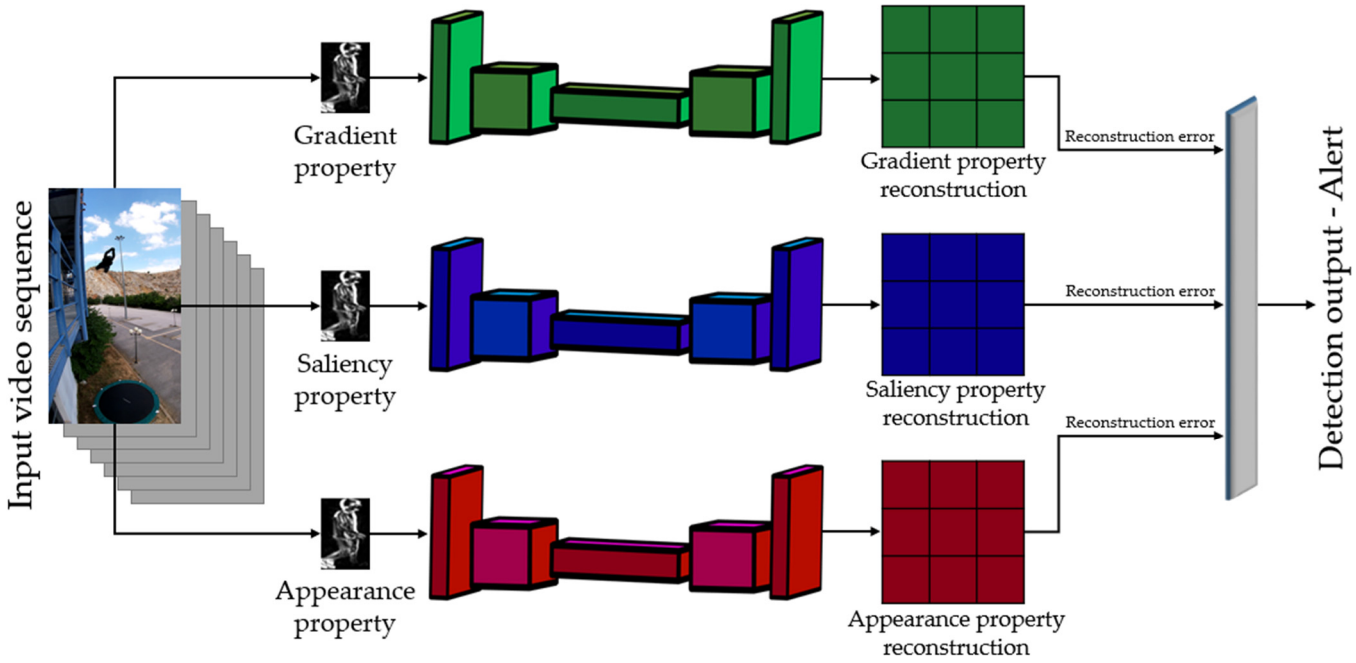


Figure 1. The overall system architecture, which utilizes a set of spatiotemporal convolutional autoencoders over multiple image properties of the RGB video sequences.

3. System Architecture

Learning Architecture

The presented system utilizes only RGB video streams to identify overboard falls. However, the simple use of raw RGB frames is not sufficient for the efficient detection of the man overboard event. To extract additional data from the visual modality, we further analyzed the camera streams to extract specific visual properties, i.e., representative vectors. To this end, the visual modality is analyzed to extract the actual frame (appearance), the gradient of the frame using a short memory window of 10 frames (movement vector), and the objectness of the current frame (saliency vector). Thus, the Appearance Property consists of the actual frame capturing. Subsequently, the Gradient Property captures the movement of the objects by taking as the input the gradient of the frame. Finally, the Saliency Property reflects how likely a window of the frame covers an object of any category. This property creates a saliency map with the same size as the frame that covers all objects of an image in a category-independent manner.

Each image property was fed into an individual spatiotemporal autoencoder. Autoencoders are a type of neural network that manages to learn efficient data encodings by training the network to ignore signal noise. Their usefulness comes from the fact that they are trained in an unsupervised manner. They are essentially composed of two main components that are trained in parallel. The dimensionality reduction component aims at extracting an efficient encoding of the input signal, while the reconstruction side tries to generate from the reduced encoding a representation as close as possible to the original input. To identify the abnormalities, the reconstruction error of each autoencoder was monitored, and when the error was bigger than the predefined threshold, an alert was raised. The selection of the threshold took place during the training to identify the exact value that maximized the detection performance. The autoencoders used for each image property had the structure presented in Figure 2. For the appearance vector, the RGB frames were converted and resized to grayscale images with a resolution of $227 \times 227 \times 1$. Lastly, it is noted that, for the analysis, minibatches of size 10 were utilized.

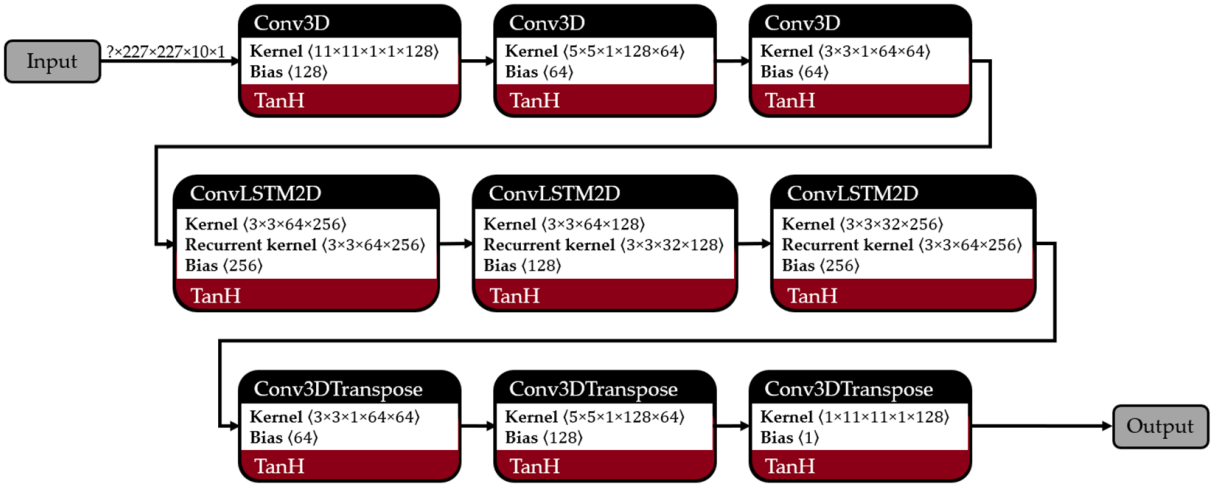


Figure 2. Structure of the individual autoencoders.

4. Experimental Evaluation

4.1. Dataset Description

To train and evaluate the proposed methodology, a mock man overboard event was conducted that concerned the fall of a human-sized dummy from the balcony of a high-rise building. In particular, the human dummy (see Figure 3), weighting 30 kg, was thrown from an approximate height of 20 m, which is roughly equivalent to two seconds of free-falling.

Figure 3. The human-sized dummy that was used during the test throws.

As depicted in Figure 4a–d, for the purposes of the specific experiment, we performed and recorded 320 test throws of the human dummy to simulate the falls and, in particular, the man overboard event (i.e., positive events). Furthermore, as presented in Figure 4e,f, we recorded numerous videos without dropping the human dummy, as well as several throws of various plastic objects, such as bottles and bags (i.e., negative events). In this way, we were able to train a deep learning network that was generalized and not prone to false-positive alerts caused by events that were not human-related.

More specifically, the experiments lasted five days and took place in the surrounding area of the Nikaia Olympic Weightlifting Hall. Since the test throws of the objects were carried out from 9:00 a.m. to 5:00 p.m. and, therefore, throughout the entire day, the acquired data varied in terms of color spaces and illumination conditions (e.g., overexposure and underexposure). In parallel, it is noted that the videos were shot under different weather conditions (e.g., cold, hot, sunny, windy, cloudy, and rainy), hence producing further variations in the general background of the falling event.

(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 4. Test throws of the dummy during the data collection experiments (as presented in [15]). The freefall (**a**–**d**) of the human dummy from different shooting angles (positive events) and various other objects, such as (**e**) plastic bags and (**f**) bottles (negative events).

Thereby, in this work, as depicted in Figure 4, we utilized a dataset that contains RGB videos featuring the (i) freefall of the human-sized dummy, (ii) normal situation, and (iii) throws of other various objects. For the data collection procedure, a GoPro Hero 7 Silver (see Figure 5) was used. It is emphasized that the acquired data are video sequences with an aspect ratio of 9:16 and, in particular, with a pixel resolution of 1080 × 1920. Moreover, the RGB sensor was set to shoot at a high frame rate and, more specifically, at 50 frames per second in order to ensure sufficient data acquisition regarding both the positive and negative events. The dataset of this paper is available online at: https://github.com/ikatsamenis/Fall-Detection/ (accessed date 14 January 2022).

(**a**) (**b**)

Figure 5. The RGB optical sensor (**a**,**b**), which was used during the data acquisition experiments to monitor the test throws of the human dummy, mounted on the perimeter of the building.

It is noted that, in order to avoid training bias, as well as to guarantee the replicability and generalization of the proposed network to other datasets, the RGB camera was placed at four different locations of the building (see Figure 4a–d). In such a way, we were able to obtain videos that varied in terms of illumination, background, distance, and shooting angle. More specifically, as illustrated in Figure 6, the sensor was placed (i) to the left of the freefall trajectory at a short distance of 7 m (see Figure 4a), (ii) to the right of the freefall trajectory at a short distance of 5 m (see Figure 4b), (iii) to the top left of the freefall trajectory at a roughly 45° angle (see Figure 4c), and (iv) to the left of the freefall

trajectory at a long distance of 13 m (see Figure 4d). Lastly, it is underlined that, to further generalize the learning process, we augmented the training data by horizontally flipping the corresponding videos.

Figure 6. The four different locations of the building (red, green, orange, and blue) where the RGB camera was placed throughout the data acquisition process regarding the freefall (yellow) of the human dummy (as presented in [15]).

4.2. Experimental Setup—Model's Training

The proposed method was implemented in the interactive environment called "Google Colaboratory", which allows the user to write Python codes through a browser. In this environment, important libraries are already installed, such as TensorFlow and Keras. This specific implementation used Python 3.7.12, Keras (1.08) and TensorFlow (2.1.0) machine learning libraries in combination with various scientific and data management libraries. The deep models were trained using a Tesla K80 GPU.

In order to train the model, a preprocessing stage was necessary. Preprocessing began with the separation of the RGB video data into the train and test sets. No falling action data were used for the train set, while falling action data were used for the test set. Subsequently, frames were exported from the video data. These frames were resized and turned into grayscale to train the proposed autoencoder model. Then, the training process was initiated by only using the data that had no falling action. Thus, these data depicted only the normal situation. On the other hand, the test data, which were used for the predictions after the training process, featured non-falling as well as falling events.

In order to study the most useful camera placement, two models were trained; the first model was designed for the horizontal view and the second one for the 45° angle view of the camera. For comparison purposes, a supervised learning method was adopted that consisted of a deep CNN classifier. In this method, the same data as in the unsupervised learning method were used, but the falling and no falling data were combined with the training process. More specifically, 60% of the entire dataset was used for the train set, 30% for the test set, and 10% for the validation set. In this method, the same preprocessing concept was followed, and the focus was on the best camera placement, as in the unsupervised learning method.

4.3. Experimental Results

The performance of the comparative networks was evaluated in the test set of the dataset described in Section 4.1. Initially, we started with a simple autoencoder over only the appearance property, tested its performance from different camera angles, and subsequently compared it with a simple CNN classifier. For this purpose, the autoencoder was trained on videos representing the normal condition, i.e., video sequences with zero numbers of falls in them. On the contrary, the CNN classifier was trained by utilizing RGB video frames that depict both positive (i.e., falls) and negative (i.e., normal situation) events. The performance evaluation of the deep learning networks took place by utilizing video frames that include the falls of the human-sized dummy, as well as an equal number of frames that depict only the normal condition. Lastly, the single autoencoder over the appearance property was expanded in such a way that it utilizes multiple image properties (i.e., appearance, gradient, and saliency) in order to enhance the detection capabilities of the proposed architecture.

It is underlined that, to guarantee adequate spatiotemporal information processing, the aforementioned algorithms were set to perform a detection every 10 frames (5 Hz). To this end, Figure 7 illustrates the freefall of the human dummy during a test throw as recorded by the RGB sensor and, in particular, the frames in which the deep models performed the fall identification task.

(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

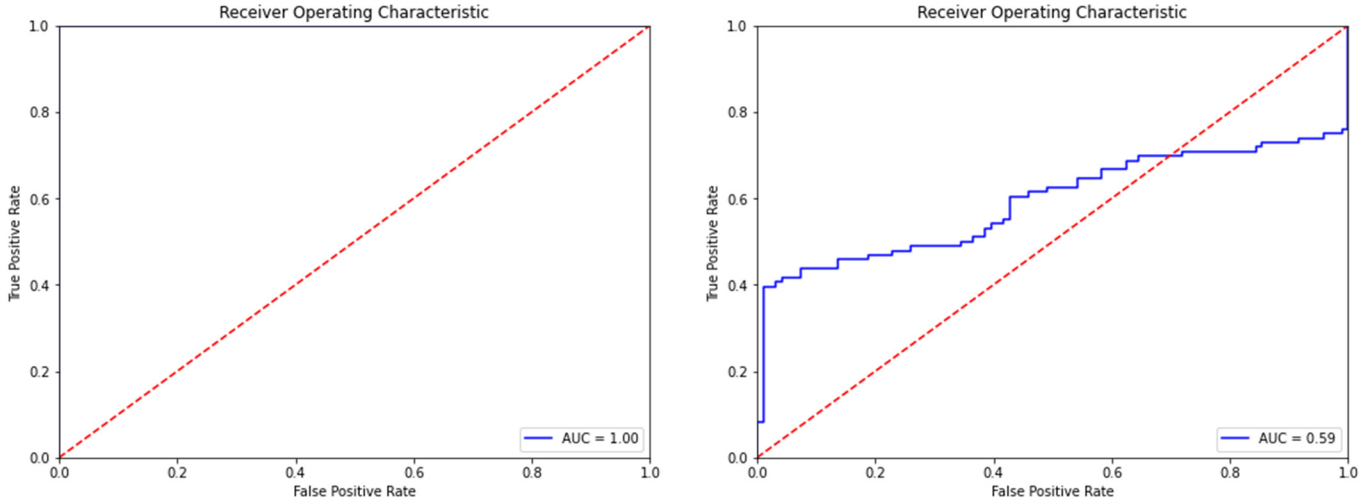
Figure 7. The freefall trajectory of the human dummy during a test throw. The deep networks were set to perform five detections per second, i.e., every 10th frame of the video sequence (**a**–**f**).

4.3.1. Performance of the Single Autoencoder with Data from Different Camera Angles

To evaluate the performance of the single autoencoder over the appearance property for RGB data that was derived from different camera angles, the Area Under the Curve (AUC) metric was employed. In particular, the AUC score is calculated in relation to the ground truth annotations at the frame level and is a common performance metric for various abnormal identification methods. In this work, it was used to measure the ability of the learning algorithm to correctly distinguish falling from no falling events and summarize the Receiver Operating Characteristics (ROC) curve of the system. The ROC curve represents the probability curve that depicts the raising of a true alert, as well as a

false alarm, and demonstrates the diagnostic ability of a deep model as its discrimination threshold is modified.

As observed in Figure 8a, the single autoencoder over the appearance property achieved an AUC score of 100% when fed with horizontal view data (see Figure 4a,b,d). On the contrary, as can be seen in Figure 8b, the algorithm demonstrated an AUC score of 59% when utilizing 45° angle view data (see Figure 4c). It is highlighted that an AUC of 100% denotes a perfect classifier, whereas an AUC of 50% corresponds to a network that produces random identification outputs. Therefore, the horizontal view model showed an excellent measure of separability. On the other hand, the 45° angle view model showed no class separation capacity whatsoever, since its AUC score was relatively close to 50%. Consequently, the AUC score proved that the horizontal view was the most suitable placement for the camera.



(**a**) (**b**)

Figure 8. The Receiver Operating Characteristics (ROC) curves of the proposed algorithm over the appearance property for the (**a**) horizontal view and (**b**) 45° angle view model.

The vast difference between the two models, in terms of their AUC scores, is mainly due to the fact that, when utilizing horizontal view RGB data, the network can more effectively learn representations from the different stages of the falling event. In particular, the vertical displacement of the object during its fall is observed in a more efficient way by placing the camera without inclination in relation to the object's trajectory (see Figure 7). On the contrary, in two consecutive frames that derive from the 45° angle view data, the vertical displacement of the object tends to be negligible. In other words, through the entire video sequence, the position of the object is almost static, and, therefore, the network fails to detect anomalies in the spatiotemporal information stream. In conclusion, the trajectory of the freefall is more efficiently depicted and analyzed by utilizing the horizontal (see Figure 4a,b,d) than the 45° angle (see Figure 4c) view of the RGB camera. Figure 8 confirms the superiority of the horizontal view model, which was eventually embedded in the proposed multi-property spatiotemporal autoencoder.

4.3.2. Performance of the Comparative Deep Learning Techniques

In parallel, the implemented comparative deep learning networks were evaluated in terms of four performance metrics, namely:

- Accuracy (*Acc*), which is the simplest of the four metrics and denotes the percentage of the correctly identified man overboard events in relation to the total amount of video sequences.

- Precision (*Prec*), which is the percentage of the correct positive detections to the total positive detections that a deep model considers. It is highlighted that a low precision score entails a high number of false alarms.
- Recall (*Rec*), which is the ratio of the correct positive detections to the total positive events in the ground truth data. It is emphasized that a low recall score implies that the model has a high number of misses.
- *F*1-score (*F*1), which is the harmonic mean of precision (*Prec*) and recall (*Rec*).

For a given set of RGB video frames, the aforementioned evaluation metrics are calculated as follows:

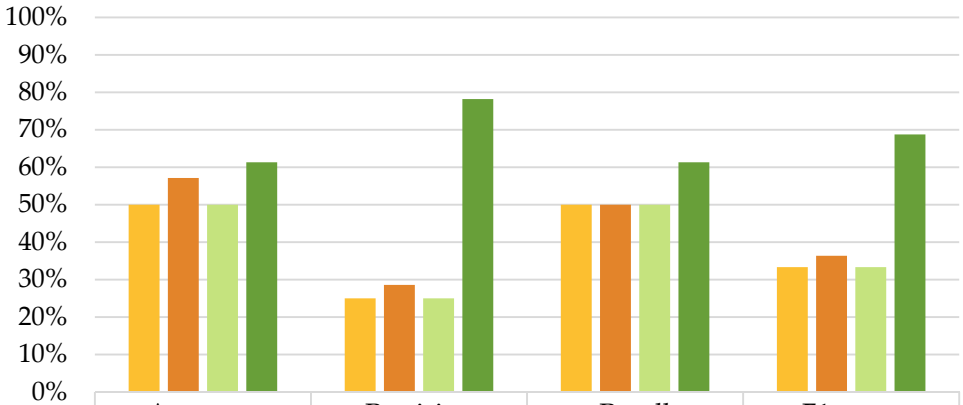
$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

$$Prec = \frac{TP}{TP + FP}, \quad (2)$$

$$Rec = \frac{TP}{TP + FN}, \quad (3)$$

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \quad (4)$$

where *TP*, *TN*, *FP*, and *FN* denote, respectively, the true-positive, true-negative, false-positive, and false-negative fall detections. In this paper, the aforementioned performance metrics were computed (in %) across all video sequences of the test set for each of the comparative deep learning techniques and can be seen in Figure 9.



	Accuracy	Precision	Recall	F1-score
Supervised paradigm (Classifier)—45° angle view	50.00%	25.00%	50.00%	33.33%
Supervised paradigm (Classifier)—Horizontal view	57.14%	28.57%	50.00%	36.36%
Unsupervised paradigm (Autoencoder)—45° angle view	50.00%	25.00%	50.00%	33.33%
Unsupervised paradigm (Autoencoder)—Horizontal view	61.35%	78.20%	61.35%	68.76%

Figure 9. Comparative analysis of the unsupervised single autoencoder vs. the supervised classifier approach for the different capturing positions of the RGB camera.

Regarding the performance of the single autoencoder, which utilizes data that derive from the 45° angle view camera, the low percentage of the metrics lies in the fact that the labeling rate is low due to the negligible vertical displacement of the human dummy in two consecutive video frames during its freefall. Thus, the unsupervised model fails to effectively identify anomalies in the spatiotemporal information stream, a fact that also

confirms its low AUC score that was presented in Section 4.3.1 and illustrated in Figure 8b. In parallel, it is clear, considering the low percentage of the metrics, that a supervised learning method, either by utilizing horizontal or 45° angle view data, is flawed and, thereby, inadequate for the purpose of this paper (see Figure 9).

On the other hand, among the comparative models, the horizontal view autoencoder demonstrated the best performance, achieving 61.35% accuracy and 68.76% *F*1-score. Additionally, the fact that its precision is significantly higher than that of the other three approaches shows that there is a negligible quantity of FP values. It is noted that precision indicates how good a deep model is when its outcome is positive. This entails that, by utilizing the horizontal view model over the appearance property, we had the minimum amount of false alarms. Furthermore, the proposed autoencoder presented the best predictive performance in terms of the recall metric. It is emphasized that recall shows how many of the positive classes the model is able to correctly predict. This implies that the horizontal view model over the appearance property showed the lowest number of misses of the critical event.

In a nutshell, concerning the placement of the camera, the horizontal view has proved to be the most suitable for the effective observation of the man overboard event, and in terms of the deep learning approach, the unsupervised autoencoder demonstrated the best identification capabilities. It is, however, highlighted that Figure 9 shows low detection rates for both supervised and unsupervised approaches. On the one hand, video queues have consecutive RGB frames, and therefore, even if the detection fails for one current frame, it is highly likely that it will succeed in the next ones. On the other hand, man overboard is a critical incident in which rapid and effective recognition plays an important role in the recovery of the victim. Thereby, we expand the horizontal view unsupervised model that showed the best results among the comparative models to improve its detection capabilities. More details on this will be given in the next subsection.

4.3.3. Performance of the Proposed Multi-Property Spatiotemporal Autoencoder

From the analysis above, it is observed that an autoencoder model analyzing streams from the horizontal view angle provides the optimal results, in terms of all four evaluation metrics (i.e., accuracy, precision, recall, and *F*1-score). These, however, still fail to achieve a performance that can be considered sufficient for using it in real-world scenarios. To this end, we mobilized an additional set of autoencoders over the several image properties, as seen in Figure 1. Based on the same annotation that was used for the comparative analysis of the autoencoder and the classifier, which was presented in Section 4.3.2 and depicted in Figure 9, we can assess the performance of the multiple autoencoder method. The performance scores achieved by the proposed multi-property spatiotemporal autoencoder on the test set of the dataset described in Section 4.1 can be seen in Figure 10.

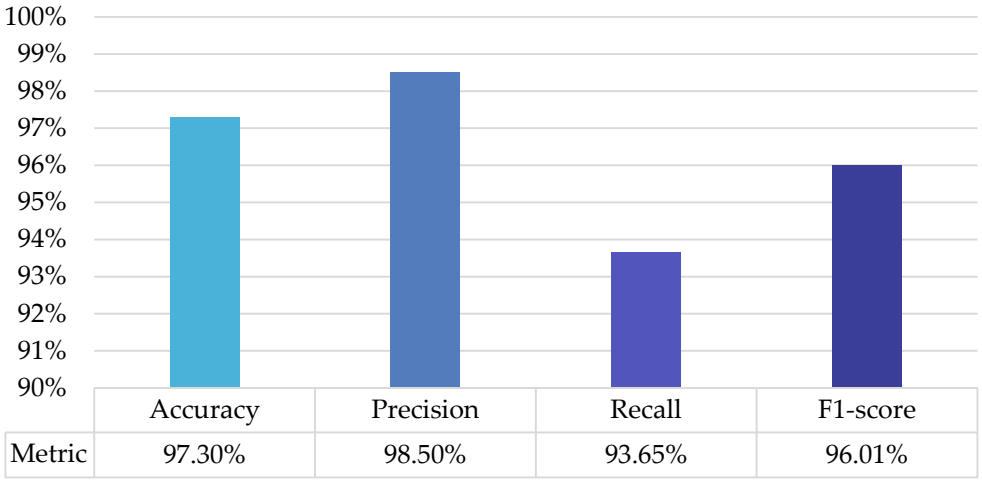


Figure 10. Performance of the unsupervised multi-autoencoder approach.

More specifically, the multi-property spatiotemporal autoencoder outperformed the single autoencoder that showed the best detection capabilities over the appearance property among the comparative deep models (see Section 4.3.2) in terms of all four performance metrics. It is noted that the single network over the appearance property presented an error rate of 38.65% and 31.24% in terms of the accuracy and $F1$-score, respectively (see Figure 9). On the other hand, the proposed model demonstrated, respectively, an error rate of 2.70% and 3.99% (see Figure 10).

To quantify this improvement in a more reliable way, we reported the relative change in the error rate in terms of accuracy and $F1$-score obtained by our proposed multi-property (i.e., appearance, gradient, and saliency) framework in relation to the error rate of the single autoencoder over only the appearance property. To this end, let us denote the relative error rate change (ε_{RC}) of a given evaluation metric as:

$$\varepsilon_{RC} = \left| \frac{\varepsilon_{mp} - \varepsilon_{sp}}{\varepsilon_{sp}} \right| \cdot 100\%, \tag{5}$$

where ε_{sp} is the error rate of the single autoencoder over the appearance property in terms of a given evaluation metric, and similarly, ε_{mp} is the error rate in terms of the same metric obtained by our proposed methodology after expanding the analysis that the autoencoder performs to multi-property (i.e., appearance, gradient, and saliency).

In particular, in terms of the accuracy metric, the multi-property spatiotemporal algorithm demonstrates an error rate improvement (see Equation (5)) of 93.01% against the traditional autoencoder over the appearance property. Similarly, the proposed network outperformed the conventional autoencoder in terms of the $F1$-score, yielding an error rate reduction (see Equation (5)) of 87.23%. In conclusion, through the proposed expansion of the spatiotemporal autoencoder, in such a way that it utilizes multiple image properties (i.e., appearance, gradient, and saliency) of the RGB video queues, the error rate was roughly reduced to 1/10 of its original value.

5. Discussion

As presented in the previous section, one can observe in the obtained evaluation scores of the proposed multi-autoencoder approach that, by analyzing several properties of the RGB data, the network's performance can be significantly increased. This is mainly due to the fact that the proposed framework analyzes data modalities that capture complementary information. To this end, the multi-property spatiotemporal autoencoder separately processes movement (gradient property) and image objectness (saliency property) in parallel with the raw visual cues (appearance property). It is noted that this is also achieved by the extraction of simple properties from the image frame and by doing only one forward propagation in the autoencoder for each image property.

Nevertheless, while this approach is efficient and can be deployed in low-power CPUs, which entails that the proposed framework can be integrated into real-time surveillance systems, it is not appropriate for usage in more complicated datasets involving large numbers of humans and/or actions. This is mainly due to the fact that each image property is analyzed separately. However, such datasets are outside of the scope of the studied scenario. For the man overboard event, the issue of false positives has mainly to do with the presence of nonhuman-caused movements (e.g., birds flying in the field of view of the camera), which has been considered in the creation of the dataset and is addressed by the utilization of the movement/gradient image property.

To this end, in the future, in order to address the aforementioned complicated scenarios, an analysis of fusing the features of all three aforementioned modalities will need to be studied. It is underlined, however, that this would require the presence of specialized learning methods that deal with data with extremely large dimensions. This happens because the fused vector of features composed by all three modalities will contain a large number of dimensions. Consequently, tensor-based learning techniques should be utilized to solve the specific curse of dimensionality.

6. Conclusions and Future Work

Identifying a man overboard event is a challenging task, since it is an incident that occurs rarely and, hence, presents a severe class imbalance problem. In this study, man overboard detection was formulated as an anomaly detection problem. We presented and evaluated an unsupervised learning algorithm for the automated recognition of such critical events, which is based on a spatiotemporal convolutional autoencoder. The employed technique models the normal conditions of the perimeter of the ship by learning the spatial and temporal features from the input video frames during the training stage and then identifies falls as abnormal behavior.

More specifically, the proposed framework uses multi-property (i.e., appearance, gradient, and saliency) analysis of RGB video streams in order to extract salient features and encodings of the normal scene utilizing a set of spatiotemporal convolutional autoencoders. Subsequently, the system can recognize a man overboard situation depending on whether the autoencoder is able or not to reconstruct a scene due to the potential existence of an abnormal event. Furthermore, to train and evaluate the performance of the proposed method, a dataset containing RGB video sequences with test throws of a human-sized dummy from the balcony of a high-rise building was demonstrated. The proposed multi-property spatiotemporal autoencoder achieved state-of-the-art results and, in particular, 97.30% accuracy and 96.01% $F1$-score on the test set of the presented dataset, surpassing other state-of-the-art approaches, such as a single autoencoder, over the appearance property and a conventional CNN classier. This entails a relative change in the error rate of 93.01% and 87.23% in terms of the accuracy and the $F1$-score, respectively. Therefore, through the proposed expansion of the autoencoder in such a way that it utilizes multiple image properties, the obtained error rate was roughly decreased to 1/10 of its original value.

Future work will concentrate on maximizing the performance of our anomaly detection scheme by utilizing additional information modalities, such as thermal imaging data and radar signals, as well as multimodal information fusion techniques for the efficient automated recognition of the man overboard event [73]. Recent studies have underlined that the use of thermal sensors is a crucial factor in various computer vision surveillance systems, since humans are warm-blooded organisms, a property that distinguishes them from their environment in thermal imagery [74]. In parallel, by leveraging radar signals the intelligent system will be able to dynamically track and monitor in real time the critical event more efficiently, thus aiding in the quick recovery of the victim [75–77]. Therefore, by fusing multi-sensor data (i.e., optical and thermal video streams, as well as radar signals), the overall performance of the automated fall detection system can be significantly improved [55]. Lastly, we will focus on additional ways for intra- and inter-property encoding of the various modalities in order to further improve the identification capabilities of the proposed maritime surveillance system.

Author Contributions: Conceptualization, N.B., I.K., A.D. and N.D.; methodology, N.B., E.E.K. and I.K.; software, N.B. and E.E.K.; validation, N.B., I.K., A.D. and N.D.; formal analysis, N.B.; investigation, N.B. and I.K.; resources, I.K.; data curation, I.K.; writing—original draft preparation, I.K. and E.E.K.; writing—review and editing, I.K. and N.B.; visualization, I.K. and N.B.; supervision, A.D. and N.D.; project administration, A.D. and I.K.; and funding acquisition, A.D. and N.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation under the call RESEARCH—CREATE—INNOVATE (project code: T1EDK-01169).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset of this work is available online at: https://github.com/ikatsamenis/Fall-Detection/ (accessed date 14 January 2022).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Örtlund, E.; Larsson, M. Man Overboard Detecting Systems Based on Wireless Technology. Bachelor Thesis, Chalmers University of Technolog, Gothenburg, Sweden, 2018.
2. Sevïn, A.; Bayilmiş, C.; Ertürk, İ.; Ekïz, H.; Karaca, A. Design and Implementation of a Man-Overboard Emergency Discovery System Based on Wireless Sensor Networks. *Turk. J. Electr. Eng. Comput. Sci.* **2016**, *24*, 762–773. [CrossRef]
3. Katsamenis, I.; Protopapadakis, E.; Voulodimos, A.; Dres, D.; Drakoulis, D. Man overboard event detection from RGB and thermal imagery: Possibilities and limitations. In Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, 30 June 2020; ACM: New York, NY, USA, 2020; pp. 1–6. [CrossRef]
4. Hennin, S.; Germana, G.; Garcia, L. Integrated Perimeter Security System. In Proceedings of the 2007 IEEE Conference on Technologies for Homeland Security, Woburn, MA, USA, 16–17 May 2007; pp. 70–75. [CrossRef]
5. Katsamenis, I.; Doulamis, N.; Doulamis, A.; Protopapadakis, E.; Voulodimos, A. Simultaneous Precise Localization and Classification of metal rust defects for robotic-driven maintenance and prefabrication using residual attention U-Net. *Autom. Constr.* **2022**, *137*, 104182. [CrossRef]
6. Katsamenis, I.; Protopapadakis, E.; Doulamis, A.; Doulamis, N.; Voulodimos, A. Pixel-level corrosion detection on metal constructions by fusion of deep learning semantic and contour segmentation. In Proceedings of the International Symposium on Visual Computing, San Diego, CA, USA, 5–7 October 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 160–169. [CrossRef]
7. Kwon, D.; Kim, H.; Kim, J.; Suh, S.C.; Kim, I.; Kim, K.J. A Survey of Deep Learning-Based Network Anomaly Detection. *Cluster Comput.* **2019**, *22*, 949–961. [CrossRef]
8. Lalos, C.; Voulodimos, A.; Doulamis, A.; Varvarigou, T. Efficient Tracking Using a Robust Motion Estimation Technique. *Multimed. Tools Appl.* **2014**, *69*, 277–292. [CrossRef]
9. Chen, Y.; Tian, Y.; He, M. Monocular Human Pose Estimation: A Survey of Deep Learning-Based Methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897. [CrossRef]
10. Rallis, I.; Georgoulas, I.; Doulamis, N.; Voulodimos, A.; Terzopoulos, P. Extraction of Key Postures from 3D Human Motion Data for Choreography Summarization. In Proceedings of the 2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games), Athens, Greece, 6–8 September 2017; pp. 94–101. [CrossRef]
11. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.H. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, 1. [CrossRef]
12. Protopapadakis, E.; Katsamenis, I.; Doulamis, A. Multi-Label Deep Learning Models for Continuous Monitoring of Road Infrastructures. In Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, 30 June 2020; ACM: New York, NY, USA, 2020; pp. 1–7. [CrossRef]
13. Feraru, V.A.; Andersen, R.E.; Boukas, E. Towards an Autonomous UAV-Based System to Assist Search and Rescue Operations in Man Overboard Incidents. In Proceedings of the 2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), Abu Dhabi, United Arab Emirates, 4–6 November 2020; pp. 57–64. [CrossRef]
14. Zhao, Y.; Yin, Y.; Gui, G. Lightweight Deep Learning Based Intelligent Edge Surveillance Techniques. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *6*, 1146–1154. [CrossRef]
15. Bakalos, N.; Katsamenis, I.; Voulodimos, A. Man Overboard: Fall Detection Using Spatiotemporal Convolutional Autoencoders in Maritime Environments. In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference, Corfu, Greece, 29 June–2 July 2021; ACM: New York, NY, USA, 2021; pp. 420–425. [CrossRef]
16. Bakalos, N.; Katsamenis, I.; Karolou, E.; Doulamis, N. Unsupervised Man Overboard Detection Using Thermal Imagery and Spatiotemporal Autoencoders. In Proceedings of the 1st International Conference on Novelties in Intelligent Digital Systems, Corfu, Greece, 30 September–1 October 2021; ACM: New York, NY, USA, 2021; pp. 256–263. [CrossRef]
17. Masci, J.; Meier, U.; Cireşan, D.; Schmidhuber, J. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In Proceedings of the Lecture Notes in Computer Science, Espoo, Finland, 14–17 June 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 52–59. [CrossRef]
18. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification. In Proceedings of the British Machine Vision Conference, Surrey, UK, 3–7 September 2012; British Machine Vision Association: Durham, UK, 2012; Volume 1, p. 12. [CrossRef]
19. Nogas, J.; Khan, S.S.; Mihailidis, A. DeepFall: Non-Invasive Fall Detection with Deep Spatio-Temporal Convolutional Autoencoders. *J. Healthc. Inform. Res.* **2020**, *4*, 50–70. [CrossRef]
20. Chowdhury, S.A.; Kowsar, M.M.S.; Deb, K. Human Detection Utilizing Adaptive Background Mixture Models and Improved Histogram of Oriented Gradients. *ICT Express.* **2018**, *4*, 216–220. [CrossRef]
21. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]

22. Zhu, Q.; Yeh, M.-C.; Cheng, K.-T.; Avidan, S. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Volume 2 (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1491–1498. [CrossRef]
23. Gajjar, V.; Khandhediya, Y.; Gurnani, A. Human Detection and Tracking for Video Surveillance: A Cognitive Science Approach. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 2805–2809. [CrossRef]
24. Mikolajczyk, K.; Schmid, C.; Zisserman, A. Human Detection Based on a Probabilistic Assembly of Robust Part Detectors. In *Lecture Notes in Computer Science, Prague, Czech Republic, 16 May 2004*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 69–82. [CrossRef]
25. Xia, L.; Chen, C.-C.; Aggarwal, J.K. Human Detection Using Depth Information by Kinect. In Proceedings of the CVPR 2011 WORKSHOPS, Colorado Springs, CO, USA, 20–25 June 2011; pp. 15–22. [CrossRef]
26. Tuzel, O.; Porikli, F.; Meer, P. Human Detection via Classification on Riemannian Manifolds. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8. [CrossRef]
27. Yang, Y.; Ramanan, D. Articulated Human Detection with Flexible Mixtures of Parts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2878–2890. [CrossRef] [PubMed]
28. Zeng, X.; Ouyang, W.; Wang, X. Multi-Stage Contextual Deep Learning for Pedestrian Detection. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 121–128. [CrossRef]
29. Ouyang, W.; Wang, X. Joint Deep Learning for Pedestrian Detection. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2056–2063. [CrossRef]
30. Voulodimos, A.S.; Kosmopoulos, D.I.; Doulamis, N.D.; Varvarigou, T.A. A Top-down Event-Driven Approach for Concurrent Activity Recognition. *Multimed. Tools Appl.* **2014**, *69*, 293–311. [CrossRef]
31. Doulamis, N.D.; Voulodimos, A.S.; Kosmopoulos, D.I.; Varvarigou, T.A. Enhanced Human Behavior Recognition Using HMM and Evaluative Rectification. In Proceedings of the First ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams—ARTEMIS '10, Firenze, Italy, 29 October 2010; ACM Press: New York, New York, USA, 2010; pp. 39–44. [CrossRef]
32. Makantasis, K.; Protopapadakis, E.; Doulamis, A.; Doulamis, N.; Matsatsinis, N. 3D Measures Exploitation for a Monocular Semi-Supervised Fall Detection System. *Multimed. Tools Appl.* **2016**, *75*, 15017–15049. [CrossRef]
33. Rougier, C.; Meunier, J.; St-Arnaud, A.; Rousseau, J. Robust Video Surveillance for Fall Detection Based on Human Shape Deformation. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 611–622. [CrossRef]
34. Yu, M.; Rhuma, A.; Naqvi, S.M.; Wang, L.; Chambers, J. A Posture Recognition Based Fall Detection System for Monitoring an Elderly Person in a Smart Home Environment. *IEEE Trans. Inf. Technol. Biomed.* **2012**, *16*, 1274–1286. [CrossRef] [PubMed]
35. Abobakr, A.; Hossny, M.; Abdelkader, H.; Nahavandi, S. RGB-D fall detection via deep residual convolutional LSTM networks. In Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, ACT, Australia, 13–18 December 2018; pp. 1–7. [CrossRef]
36. Adhikari, K.; Bouchachia, H.; Nait-Charif, H. Activity recognition for indoor fall detection using convolutional neural network. In Proceedings of the 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), Nagoya, Japan, 8–12 May 2017; pp. 81–84. [CrossRef]
37. Cameiro, S.A.; da Silva, G.P.; Leite, G.V.; Moreno, R.; Guimaraes, S.J.F.; Pedrini, H. Multi-stream deep convolutional network using High-Level features applied to fall detection in video sequences. In Proceedings of the 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), Osijek, Croatia, 5–7 June 2019; pp. 293–298. [CrossRef]
38. Espinosa, R.; Ponce, H.; Gutiérrez, S.; Martínez-Villaseñor, L.; Brieva, J.; Moya-Albor, E. A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the UP-Fall detection dataset. *Comput. Biol. Med.* **2019**, *115*, 103520. [CrossRef] [PubMed]
39. Ge, C.; Gu, I.Y.-H.; Yang, J. Co-saliency-enhanced deep recurrent convolutional networks for human fall detection in E-healthcare. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. In Proceedings of the IEEE Engineering in Medicine and Biology Society, Annual International Conference, Honolulu, HI, USA, 18–21 July 2018; pp. 1572–1575. [CrossRef]
40. Hsieh, Y.-Z.; Jeng, Y.-L. Development of home intelligent fall detection IoT system based on feedback optical flow convolutional neural network. *IEEE Access Pract. Innov. Open Solut.* **2017**, *6*, 6048–6057. [CrossRef]
41. Hwang, S.; Ahn, D.; Park, H.; Park, T. Poster abstract: Maximizing accuracy of fall detection and alert systems based on 3D convolutional neural network. In Proceedings of the 2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI), Pittsburgh, PA, USA, 18–21 April 2017; pp. 343–344.
42. Kasturi, S.; Filonenko, A.; Jo, K.-H. Human Fall Recognition using the Spatiotemporal 3D CNN. In Proceedings of the IW-FCV2018, Hakodate, Japan, 21–23 February 2018; pp. 1–3.
43. Li, X.; Pang, T.; Liu, W.; Wang, T. Fall detection for elderly person care using convolutional neural networks. In Proceedings of the 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 14–16 October 2017; pp. 1–6. [CrossRef]
44. Li, S.; Xiong, H.; Diao, X. Pre-impact fall detection using 3D convolutional neural network. In Proceedings of the 2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR), Toronto, ON, Canada, 24–28 June 2019; pp. 1173–1178. [CrossRef]

45. Lie, W.-N.; Le, A.T.; Lin, G.-H. Human fall-down event detection based on 2D skeletons and deep learning approach. In Proceedings of the 2018 International Workshop on Advanced Image Technology (IWAIT), Chiang Mai, Thailand, 7–9 January 2018; pp. 1–4. [CrossRef]
46. Lin, H.Y.; Hsueh, Y.L.; Lie, W.N. Convolutional recurrent neural networks for posture analysis in fall detection. *J. Inf. Sci. Eng.* **2018**, *34*, 577–591. [CrossRef]
47. Lu, N.; Wu, Y.; Feng, L.; Song, J. Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 314–323. [CrossRef] [PubMed]
48. Lu, N.; Ren, X.; Song, J.; Wu, Y. Visual guided deep learning scheme for fall detection. In Proceedings of the 2017 13th IEEE Conference on Automation Science and Engineering (CASE), Xi'an, China, 20–23 August 2017; pp. 801–806. [CrossRef]
49. Rahnemoonfar, M.; Alkittawi, H. Spatio-temporal convolutional neural network for elderly fall detection in depth video cameras. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 2868–2873. [CrossRef]
50. Shen, L.; Zhang, O.; Cao, G.; Xu, H. Fall detection system based on deep learning and image processing in cloud environment. In *Conference on Complex, Intelligent, and Software Intensive Systems, Kunibiki Messe, Matsue, Japan, 4–6 July 2018*; Springer International Publishing: Cham, Switzerland, 2018; pp. 590–598. [CrossRef]
51. Tao, X.; Yun, Z. Fall prediction based on biomechanics equilibrium using Kinect. *Int. J. Distrib. Sens. Netw.* **2017**, *13*. [CrossRef]
52. Rougier, C.; Meunier, J. Fall detection using 3d head trajectory extracted from a single camera video sequence. In Proceedings of the First International Workshop on Video Processing for Security (VP4S-06), Quebec City, QC, Canada, 7–9 June 2006; pp. 7–9.
53. Tsai, T.-H.; Hsu, C.-W. Implementation of fall detection system based on 3D skeleton for deep learning technique. *IEEE Access Pract. Innov. Open Solut.* **2019**, *7*, 153049–153059. [CrossRef]
54. Zhou, J.; Komuro, T. Recognizing fall actions from videos using reconstruction error of variational autoencoder. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3372–3376. [CrossRef]
55. Zhou, X.; Qian, L.-C.; You, P.-J.; Ding, Z.-G.; Han, Y.-Q. Fall detection using convolutional neural network with multi-sensor fusion. In Proceedings of the 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), San Diego, CA, USA, 23–27 July 2018; pp. 1–5. [CrossRef]
56. Del Giorno, A.; Bagnell, J.A.; Hebert, M. A Discriminative Framework for Anomaly Detection in Large Videos. In *Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 334–349. [CrossRef]
57. Dutta, J.; Banerjee, B. Online Detection of Abnormal Events Using Incremental Coding Length. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; AAAI Press: Palo Alto, CA, USA, 2015; Volume 29, pp. 3755–3761. [CrossRef]
58. Ionescu, R.T.; Smeureanu, S.; Alexe, B.; Popescu, M. Unmasking the Abnormal Events in Video. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2914–2922. [CrossRef]
59. Mo, X.; Monga, V.; Bala, R.; Fan, Z. Adaptive Sparse Representations for Video Anomaly Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *24*, 631–645. [CrossRef]
60. Jiang, F.; Wu, Y.; Katsaggelos, A.K. A Dynamic Hierarchical Clustering Method for Trajectory-Based Unusual Video Event Detection. *IEEE Trans. Image Process.* **2009**, *18*, 907–913. [CrossRef] [PubMed]
61. Ribeiro, M.; Lazzaretti, A.E.; Lopes, H.S. A Study of Deep Convolutional Auto-Encoders for Anomaly Detection in Videos. *Pattern Recognit. Lett.* **2018**, *105*, 13–22. [CrossRef]
62. Chalapathy, R.; Menon, A.K.; Chawla, S. Robust, Deep and Inductive Anomaly Detection. In *Machine Learning and Knowledge Discovery in Databases, Skopje, Macedonia, 18–22 September 2017*; Springer International Publishing: Cham, Switzerland, 2017; pp. 36–51. [CrossRef]
63. Gutoski, M.; Aquino, N.M.R.; Ribeiro, M.; Lazzaretti, A.E.; Lopes, H.S. Detection of Video Anomalies Using Convolutional Autoencoders and One-Class Support Vector Machines. In Proceedings of the XIII Brazilian Congress on Computational Intelligence, Rio de Janeiro, Brazil, 30 October–1 November 2017. [CrossRef]
64. Tran, H.; Hogg, D. Anomaly Detection Using a Convolutional Winner-Take-All Autoencoder. In Proceedings of the British Machine Vision Conference 2017, London, UK, 4–7 September 2017; British Machine Vision Association: Durham, UK, 2017.
65. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning Temporal Regularity in Video Sequences. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 30 June 2016; pp. 733–742. [CrossRef]
66. Munawar, A.; Vinayavekhin, P.; De Magistris, G. Spatio-Temporal Anomaly Detection for Industrial Robots through Prediction in Unsupervised Feature Space. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 1017–1025. [CrossRef]
67. Xu, D.; Yan, Y.; Ricci, E.; Sebe, N. Detecting Anomalous Events in Videos by Learning Deep Representations of Appearance and Motion. *Comput. Vis. Image Underst.* **2017**, *156*, 117–127. [CrossRef]
68. Zhao, Y.; Deng, B.; Shen, C.; Liu, Y.; Lu, H.; Hua, X.-S. Spatio-Temporal AutoEncoder for Video Anomaly Detection. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; ACM: New York, NY, USA, 2017; pp. 1933–1941. [CrossRef]

69. Rezvanian, A.R.; Imani, M.; Ghassemian, H. Patch-Based Sparse and Convolutional Autoencoders for Anomaly Detection in Hyperspectral Images. In Proceedings of the 2020 28th Iranian Conference on Electrical Engineering (ICEE), Tabriz, Iran, 4–6 August 2020; pp. 1–5. [CrossRef]
70. Sabokrou, M.; Fayyaz, M.; Fathy, M.; Klette, R. Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes. *IEEE Trans. Image Process.* **2017**, *26*, 1992–2004. [CrossRef] [PubMed]
71. Wang, L.; Zhou, F.; Li, Z.; Zuo, W.; Tan, H. Abnormal Event Detection in Videos Using Hybrid Spatio-Temporal Autoencoder. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), New York, NY, USA, 7–10 October 2017; pp. 2276–2280. [CrossRef]
72. Xu, D.; Ricci, E.; Yan, Y.; Song, J.; Sebe, N. Learning Deep Representations of Appearance and Motion for Anomalous Event Detection. In Proceedings of the British Machine Vision Conference 2015, Swansea, UK, 7–10 September 2015; British Machine Vision Association: Durham, UK, 2015. [CrossRef]
73. Bakalos, N.; Voulodimos, A.; Doulamis, N.; Doulamis, A.; Ostfeld, A.; Salomons, E.; Caubet, J.; Jimenez, V.; Li, P. Protecting Water Infrastructure from Cyber and Physical Threats: Using Multimodal Data Fusion and Adaptive Deep Learning to Monitor Critical Systems. *IEEE Signal Process. Mag.* **2019**, *36*, 36–48. [CrossRef]
74. Mehta, V.; Dhall, A.; Pal, S.; Khan, S.S. Motion and Region Aware Adversarial Learning for Fall Detection with Thermal Imaging. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 6321–6328. [CrossRef]
75. Sheu, B.-H.; Yang, T.-C.; Yang, T.-M.; Huang, C.-I.; Chen, W.-P. Real-Time Alarm, Dynamic GPS Tracking, and Monitoring System for Man Overboard. *Sens. Mater.* **2020**, *32*, 197–221. [CrossRef]
76. Tsekenis, V.; Armeniakos, C.K.; Nikolaidis, V.; Bithas, P.S.; Kanatas, A.G. Machine Learning-Assisted Man Overboard Detection Using Radars. *Electronics* **2021**, *10*, 1345. [CrossRef]
77. Armeniakos, C.K.; Nikolaidis, V.; Tsekenis, V.; Maliatsos, K.; Bithas, P.S.; Kanatas, A.G. Human fall detection using mmWave radars: A cluster-assisted experimental approach. *J. Ambient. Intell. Humaniz. Comput.* **2022**, 1–13. [CrossRef]