applied sciences

MDPI

Article

Novel Hand Gesture Alert System

Sebastien Mambou [1], Ondrej Krejcar [1,*], Petra Maresova [2], Ali Selamat [1,3,4,5] and Kamil Kuca [1,3]

1 Center for Basic and Applied Research, Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic
2 Department of Economy, Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic
3 Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, Kuala Lumpur 54100, Malaysia
4 Media and Games Center of Excellence (MagicX), Universiti Teknologi Malaysia, Skudai 81310, Malaysia
5 School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), Skudai 81310, Malaysia
* Correspondence: ondrej.krejcar@uhk.cz

Received: 7 August 2019; Accepted: 15 August 2019; Published: 19 August 2019

Abstract: Sexual assault can cause great societal damage, with negative socio-economic, mental, sexual, physical and reproductive consequences. According to the Eurostat, the number of crimes increased in the European Union between 2008 and 2016. However, despite the increase in security tools such as cameras, it is usually difficult to know if an individual is subject to an assault based on his or her posture. Hand gestures are seen by many as the natural means of nonverbal communication when interacting with a computer, and a considerable amount of research has been performed. In addition, the identifiable hand placement characteristics provided by modern inexpensive commercial depth cameras can be used in a variety of gesture recognition-based systems, particularly for human-machine interactions. This paper introduces a novel gesture alert system that uses a combination of Convolution Neural Networks (CNNs). The overall system can be subdivided into three main parts: firstly, the human detection in the image using a pretrained "You Only Look Once (YOLO)" method, which extracts the related bounding boxes containing his/her hands; secondly, the gesture detection/classification stage, which processes the bounding box images; and thirdly, we introduced a module called "counterGesture", which triggers the alert.

Keywords: sexual assault; CNN; gesture-recognition-based systems

1. Introduction

In a sexual assault, the assailant assaults the victim quickly and brutally, without any prior contact, usually at night in a public place. This can be done by physical force or threats of force, or by the abuser giving the victim alcohol as part of the crime. Sexual assault includes rape and sexual coercion [1]. In the United States, a significant number of women face sexual assault every day, and about one in three women have been victims of this crime [2]. Similarly, the number of sexual assaults in Europe increased between 2008 and 2016 [3] despite the tremendous increase in security tools such as cameras, which require special human attention to analyze the scene [4–6]. Most of the studies proposed in the past few years tend to answer the question: was it a sexual assault? However, few of them focus on the early detection of this crime [7,8]. Indeed, this is a difficult task, as the posture of both individuals (the rapist and the victim) might be both on the same abscissa of the camera, as shown in Figure 1. Regardless of the victim's position, a part of his/her body may be visible most of the time, especially the hands, which may describe a specific pattern if the victim is aware of what is known as the "security-gestures" described in this article. This article aims to present a new hand gesture

alert system, which takes advantage of a defined set of gestures that can trigger a warning when the described computer vision system detects them.

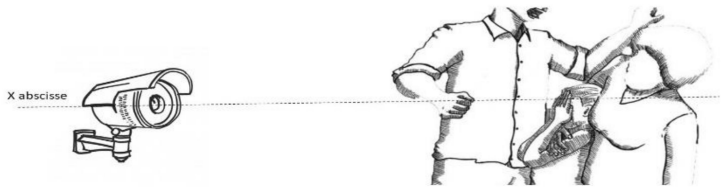

Figure 1. The body of the aggressor prevents the ability to see the entire body of the victim.

To achieve this aim, the system uses human detection, hand extraction, and Convolution Neural Networks (CNNs) [9–12]. It detects human bodies in a video, extracts the region of interest (hands) and detects the hand gesture that will be processed and will trigger the alert, if the hand gesture corresponds to one of the predefined hand gestures. The remaining part of this article is structured as follows: first of all, the related work will be presented; secondly, the proposed architecture will be addressed; thirdly, experiments, results, and discussion will be presented; lastly, a summary of the proposed work and further perspectives are offered in the conclusion.

2. Related Work

Several researchers focus on human detection and surveillance; in addition, the detection of gestures is attracting increasing interest, as is the classification of gestures. The following section gives the reader an overview of the state-of-the-art work in these areas.

2.1. Human Detection

In today's frameworks, the extraction of Regions of Interest (ROIs) and the representation of characteristics are the two main factors under study [13–18]. In [19], the difference in intensity of an individual pixel is incorporated into shape-oriented features to capture salient features. However, the framework has selected significant thresholds based on assumptions. In [20], the foregrounds are separated by subtracting the background and are then sorted by a Support Vector Machine (SVM). However, the frame only detects the upper part of the human body, and Histogram of Oriented Gradients (HOG) [21] features [22] are combined into a composite local feature. Nevertheless, the extended dimension of the composite function increases the processing costs of the system, and the other commonly used feature descriptors are shapelet, Edge Orientation Histogram (EOH) feature and Haar wavelet function. In the case of partial occlusion, it is more effective to partially detect the human body [23] rather than the whole body. However, improved accuracy also increases treatment costs. The system proposed in [24] requires about several seconds to process a single frame.

Researchers are also interested in setting up a tracking algorithm using RGB (Red Green Blue) video streams. A tracking algorithm typically consists of two different components: a local component, which includes the characteristics extracted from the target being tracked; and global characteristics, which determine the probability of locating the target. The use of a CNN to extract the characteristics of the target being tracked is a prevalent and effective method. This approach focuses primarily on object detection and local features of the target to be used for monitoring purposes [6–9]. Qi et al. [23] used two different CNN structures to distinguish the target from other distractors in the scene. Wang et al. [24] devised a structure composed of two distinct parts: a shared part shared by all the training videos and a multi-domain part that classified different videos in the training set. The first part extracted the common features to be used for tracking. Fu et al. [25] designed a CNN-based discriminant filter to obtain local characteristics.

2.2. Hand Gesture Recognition (Detection and Classification)

Looking at the previous work, we can see that various studies are dealing differently with the classification and location issues related to gestures. Regarding the location of the hand, also known as the hand detection, the authors of [26–28] extract the hand from the body using depth indices and by setting a threshold, estimated at a specific moment. The authors of [29,30] used skin color maps, and the authors of [31,32] achieved better segmentation results using both depth thresholding and skin detection (using color). In terms of classification, several CNN-based approaches relied on hand-crafted features [33,34], which can capture information about the silhouette, shape, and structure.

In [35], the authors presented a 3D dynamic system that helps to recognize gestures using hand pose information. More precisely, the authors used the natural structure of the topology of the hand—called the skeletal data of the hand—to extract kinematic descriptors from the actual hand of the sequence of gestures. Using a Fisher kernel and a multilevel temporal pyramid, respectively, the descriptors were encoded in a temporal and statistical representation. Considering a feature vector calculated over the entire pre-segmented gesture, an improvement in the recognition can be achieved by associating a linear SVM classifier directly at the end (see Figure 2).

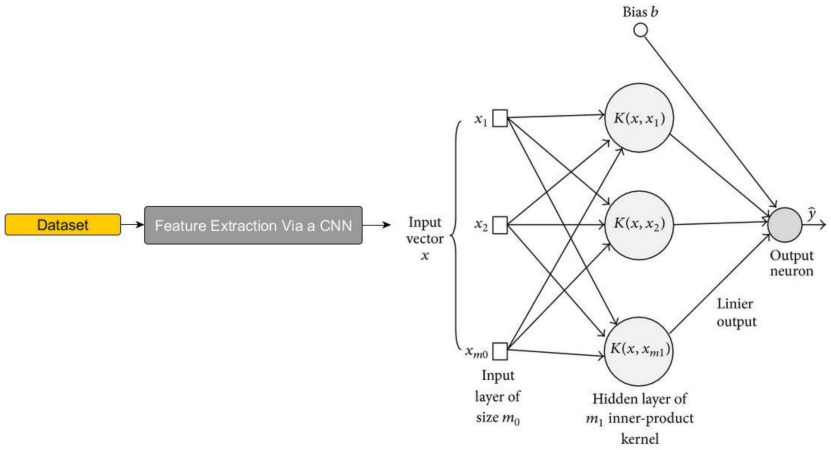

Figure 2. Improvement in the recognition by adding a Support Vector Machine (SVM) as the final component.

In terms of social factors, while browsing the literature, the need to acquire statistical data appeared. Indeed, many databases are available in the United States and offer real visibility on the increasing crime rate in general and in relation to sexual assault. Based on the data collection in [33], Figure 3 presents a forecast that reveals the areas with a high risk of experiencing a significant amount of sexual assault in the United States over the next three years.

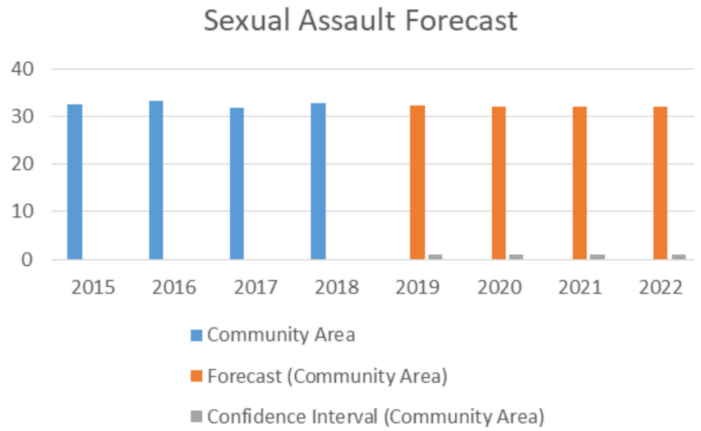

Figure 3. An estimation of future sexual assault cases in the USA [33].

3. Proposed Model

In this part, the alert system based on a hierarchical convolution neural network is described. Its architecture is composed of three main parts: the extractions of the related bounding boxes containing subject hands, the gesture detection/classification stage, and the counterGesture. After the presentation of the architecture, we will describe each component and move to the experiment section.

For the architecture, we envision an environment where many people walk on the streets. The system below takes as its input the video stream of the scene and processes it to identify if a subject is facing sexual assault. To do this, the video stream will be segmented into several frames that will be processed continuously. Based on Figure 4, we can distinguish the extraction of the regions of interest, which are the hands of everyone present in the scene. It is essential to specify that each hand is associated with its owner so that we know which subject triggers the alert.

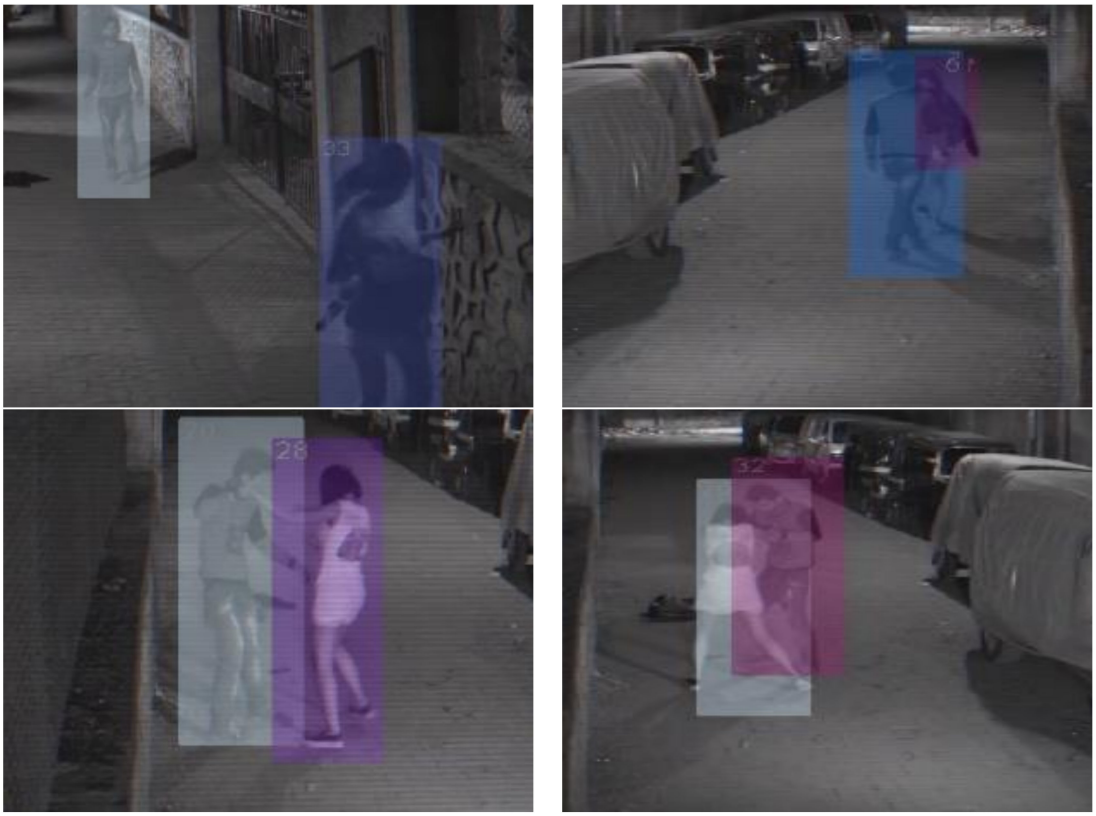

Figure 4. Multiple object tracking and extraction of the regions of interest (ROIs) or tracks. In these images, pedestrians are detected, and an identification number is assigned to each track.

3.1. Human Detector and Regions of Interest (ROIs)

Researchers are addressing the problem of multiple object tracking (MOT) with neural networks. They are doing this primarily by building robust models that capture information about movement, appearance, and interactions between objects. Considering the issue of MOT, we adopt a conventional methodology to follow different hypotheses with Kalman recursive filtering and image-by-image association. To further illustrate our idea, consider the following situation: when an object is obstructed for a more extended period, the following Kalman filter predictions increase the uncertainty associated with the position of the object (see Figure 4). We use a standard Kalman filter with a constant velocity motion and a linear observation model, where we take the delimitation coordinates (u, v, and h) as the direct observation of the state of the object.

For each track k (bounding boxes associated to the same identifier ID), we count the number of frames since the last successful measurement association a_k. This counter is incremented when

previewing the Kalman filter and reset to 0 when the track has been associated with a measure. Besides, Algorithm 1 describes how bounding boxes associated to IDs are processed in the human detector module. Furthermore, tracks that exceed a predefined maximum age A_{max} are considered to have left the scene and are removed from the set of tracks. New track assumptions are initiated for each detection that cannot be associated with an existing track. These new tracks are classified as tentative during their first three frames. In the meantime, we expect a successful metric association at each time step. Tracks that are not successfully associated with the measure in their first three frames are deleted.

As a result, the mass probability propagates in the state space, and the probability of observation decreases. Intuitively, the association metric should take into account this dispersion of the probability mass by increasing the measurement distance of the track. Counter-intuitively, when two tracks compete for the same detection, the Mahalanobis calculation should be used:

$$\theta_{i,j}^{1} = \left(d_j - y_i\right)^T S_i^{-1} \left(d_j - y_i\right)^T \tag{1}$$

where we note the projection of the i^{th} track distribution in the measuring space by (y_i ; S_i) and the j-th terminal box detection by d_j .

Distance promotes more considerable uncertainty because it effectively reduces the standard deviation distance of any detection relative to the projected runway average.

This behavior is undesirable because it can lead to increased fragmentation of unstable tracks and tracks in general. Therefore, we used a pairing cascade that gives priority to the most frequently seen objects in order to encode our notion of distributed probability in the association probability.

The human detector presented in this article is based on the "You Only Look Once (YOLO)" [36–39] method, which discretizes the output space of selected images into a set of default images of different formats and scales per map location. At the time of the prediction, the network generates scores for the presence of each object category in each default zone and produces adjustments to the area to better match the shape of the object. Also, the network combines the predictions of several feature maps with different resolutions to handle objects of various sizes naturally. The following human detector algorithm can be given:

Algorithm 1: Human detector.

Input: max age (A_{max}), *Indices*
I = {1, … , N}, *indices per Detection* (I_D) = {$n|n \in \mathbb{N}$}
Output: match $\mathcal{M}c$ *and unmatch*
Cost Co = [$Co_{i,j}$] ← $\theta d_{i,j} + (1 - \theta) d_{i,j}^2$
Gate $G \leftarrow \prod_{m=1}^{2} b_{i,j}^{m}$
$\mathcal{M} \leftarrow \{\}$, $\leftarrow I_D$

For z = 1 **to** A_{max}, **do**
  $I_z \leftarrow \{i \in I | a_i = n\}$
  $x_{i,j} \leftarrow Min_{cost}$ (Co, I,)
  $\mathcal{M}c \leftarrow \mathcal{M}c \cup \left\{\{i, j\} \middle| b_{i,j}.x_{i,j} > 0\right\}$
  $\leftarrow \cup \left\{\{i, j\} \middle| \sum b_{i,j}.x_{i,j} > 0\right\}$
End

Our pre-trained human detector YOLOv3 [36,40,41] (Table 1) is configured to obtain for almost everyone (human) in the image (I_i) a bounding box that surrounds him. Considering a video V subdivided into n frames as

$$V = \sum_{i=1}^{n} I_i \, , \tag{2}$$

each I_i frame might contain multiple subjects present in I_{i+1} frame. It will be interesting to keep track of all these subjects, so we use a so-called object tracking, which uses multiple detections to identify a specific object over time. To address this requirement, the model uses an easy and fast algorithm called SORT (Simple Online and Real-time Tracking) [42], which obtains references for the objects in the image. Therefore, instead of the regular detections, which include the coordinates of the bounding box and a class prediction, we obtain tracked subjects. These also include an object ID, which is associated with the ROI, so that for each frame, we know which hand belongs to which subject in the frame.

Table 1. The human detector takes advantage of the YOLOv3 model, which extracts the related bounding boxes containing the ROI (hands).

Layers	Filters	Stride	Output
Conv	32	3×3	224×224
Maxpool		$\frac{1}{2} \times 2 \times 2$	112×112
Conv	64	3×3	112×112
Maxpool		$\frac{1}{2} \times 2 \times 2$	28×28
Conv	128	3×3	28×28
Conv	64	1×1	28×28
Conv	128	3×3	28×28
Maxpool		$\frac{1}{2} \times 2 \times 2$	28×28
Conv	256	3×3	28×28
Conv	128	1×1	28×28
Conv	256	3×3	28×28
Maxpool		$\frac{1}{2} \times 2 \times 2$	14×14
Conv	512	3×3	14×14
Conv	256	1×1	14×14
Conv	512	3×3	14×14
Conv	256	1×1	14×14
Conv	512	3×3	14×14
Maxpool		$\frac{1}{2} \times 2 \times 2$	7×7
Conv	1024	3×3	7×7
Conv	512	1×1	7×7
Conv	1024	3×3	7×7
Conv	512	1×1	7×7
Conv	1024	3×3	7×7
Conv	1000	1×1	7×7
AveragPool		Global	1000
Softmax			

Each bounding box area A_{F_i} of the hands is further extracted as an individual frame F_i, so that

$$G = \sum_{i=1}^{n} A_{F_i} \ll A_{I_i}\,, \tag{3}$$

which will be concatenated (time-wise, t) to those of the same subject from the primary image (I). The simulation of one video recording for each submitted video will be combined with another individual video as input V_c for the next module, as:

$$V_c = \sum_{j=1}^{m} \sum_{i=1}^{n} A_{F_{ij}} \; or \; \sum_{j=1}^{m} G_j \tag{4}$$

3.2. Detector and Classifier

Over the past few years, CNN-based models have shown impressive results when they are performing gesture and action recognition tasks. CNN 3D architectures are distinguished mainly by

video analysis because they use temporal relationships between images. A new frame is described in the following section, whose goal is the detection and recognition of a specific hand gesture that will trigger the alarm.

Detector: Since we have no limitations regarding the size of the model, another architecture, with excellent classification performance, can be selected by a classifier. This leads us to use two recent 3D CNN architectures [43], the previously described YOLOv3 as the detector and a custom 3D CNN with a novelty, the introduction of "counterGesture", which is activated once a gesture is classified as belonging to our specific set. Given the context of this paper, the detector can be described as a tool responsible for processing sequential frames (video) and activating the classifier if there is a potential gesture in the video. It is worth mentioning that the human detector module mentioned above will have as its output a collection of hand images describing a gesture and associated to each subject. An algorithm (Algorithm 2) is given below.

Algorithm 2: Gesture Recognition

Input: V_c
Output: Specific Gesture
For j = 1 **to** m, **do**
  For each "frame window" G_j, **do**
    Process a batch of hand images
    classifierIsActivated $\leftarrow True$
    alpha $\leftarrow Probability_{(i-1)} \times (i-1)$
    meanProbaility $= \left(alpha + G_j \times probability_i\right)/i$
    $(max_1, max_2) = \underset{gesture}{max}$ [meanProbaility]$_2$
    If $(max_1 - max_2) \geq t_{early}$, **then**
      isEarlyDetect $\leftarrow$ "$True$"
      Return gesture (max_1)
    i $\leftarrow i+1$
  $j \leftarrow j+1$

Classifier: Depending on the proposed model, any classifier with good accuracy can be used. As shown in Table 2, we have listed the parameters of the classifier used in this article, which is a 3D convolutional neural network. Besides, we designed the model so that the number of parameters, P(3D), is greater than the number of parameters, P(2D), of a conventional 2D convolutional neural network. It should be mentioned that 3D CNNs require more training data to avoid overfitting. For this reason, we first trained our classifier on a well-known dataset, "Jester" [24], which is the largest hand gesture dataset (public dataset); then, the model was fine-tuned on nvGesture datasets with direct consequences, namely, accuracy and training time.

Table 2. The model parameters of the proposed 3D Convolution Neural Network (CNN) classifier.

Layers	Filters	Stride	Output
Conv 3D	84	$(1,2,2)$	64×64
Maxpool 3D		$(1,2,2), 64 \times 64$	
Conv 3D	64	$(2,2,2)$	128×128
Maxpool 3D		$(2,2,2), 128 \times 128$	
Conv 3D	128	$(2,2,2)$	256×256
Maxpool 3D		$(2,2,2), 256 \times 256$	
Conv 3D	128	$(2,2,2)$	256×256
Maxpool 3D		$(2,2,2), 256 \times 256$	
FC		$(12800, 512)$	
FC		$(512,\ 25)$	

CounterGesture: the system comes with a novel module, which describes the counterGesture as a module responsible for counting the occurrences of gestures similar to the predefined set of gestures (Figure 5). As mentioned in the description of Figure 6, the counterGesture comes with two operators in its architecture: the first one sets the flag to 1, and the listener 1 starts the timer; the second operator is responsible for checking the time spent already and the value contained in the incrementor, and for triggering the alert if the condition below is respected.

$$\text{getTime} < n \text{ and Counter} == 3 \tag{5}$$

where n represents a predefined maximum duration allowed to collect the three gestures.

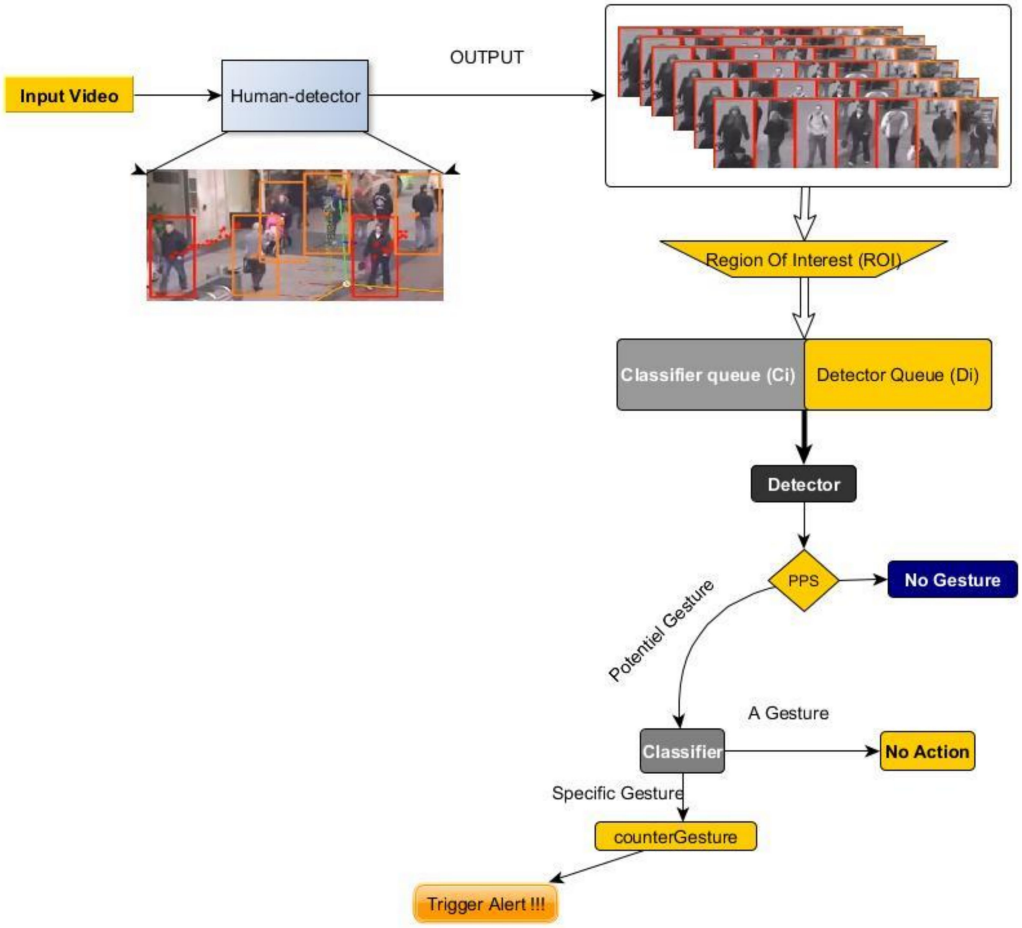

Figure 5. A video is submitted to a module called the human detector, which extracts images containing a human. These frames are processed using scrolling sliding windows in which the detection queue is placed at the very beginning of the classifier assignment queue. If the detector recognizes an action/gesture, the classifier is activated via the Post Processing Service (PPS) [44] and, if it corresponds to one of the gestures contained in our specific set, the alert is triggered.

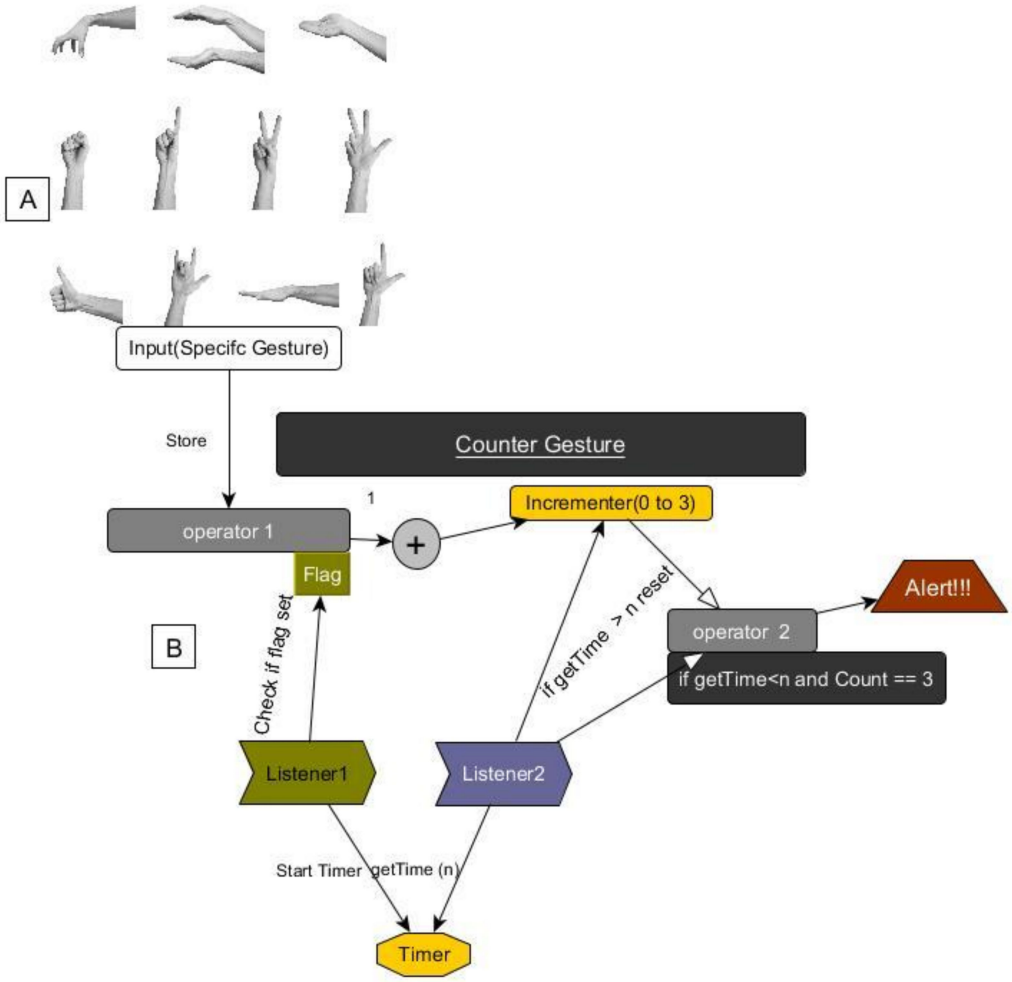

Figure 6. (**A**) Set of predefined images that the counterGesture considers. (**B**) In case the classifier identifies one of these gestures, the counterGesture will increment its counter up to 3 and trigger the alert, if there are at least three occurrences of the predefined gestures.

4. Experiments

The overall system can be divided into three parts: first, the extraction of the ROI as an output of the human detection module; second, the image classification, which categorizes all of the hand frames into one possible gesture; and third, a match that will trigger the alert. During the training part, we trained these three components separately. Furthermore, the counterGesture counts only the number of correspondences between the predicted gesture and an element of the predefined set of gestures. The latter is shown in Figure 6.

The EgoGesture dataset is a new multimodal dataset for the egocentric recognition of hand gestures [45], and it was created not only for the detection of gestures in continuous data but also for the classification of segmented gestures. This dataset contains eighty-three classes of dynamic and static gestures collected from six outdoor and indoor scenes. We organized the training set, validation and test set by separating topics with a 3:1:1 ratio, which gave respectively: 1239 elements, 411 validation videos, and 431 test videos, with 14,416, 4768, and 4977 gesture samples. All models were first pre-trained on the Jester dataset [45]. For test set evaluations, we used both the training and the validation games for the training.

To perform our experiments, we used a specific device as described in the table (Table 3). We were able to extract the ROI (Figures 7 and 8) and classify the gesture (Figure 7). In addition, Figure 9 shows the accuracy of our classifier after 350 epochs.

Table 3. Characteristics of the testing environment containing a graphic card "GeForce GTX 1070".

Station Characteristics	
CPU	Intel(R) Core ™ i7-3.20 GHz
RAM	4096 MB × 2
Device 0: "GeForce GTX 1070"	
Components	Specifications
CUDA Driver Version/ Runtime Version	10.1/10.1
CUDA Capability Major/Minor Version Number	6.1
Total Amount of Global Memory	8118 MB
(15) Multiprocessors (MP), (128) CUDA Cores/MP	1920 CUDA Cores
GPU Max Clock Rate:	1721 MHz
Memory Bus Width	256 bits

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

Figure 7. YouTube video containing a sexual assault attempt [46]. (**a**) We can see the victim being assault by an individual. (**b**) Our system performs an analysis of the scene, extracts the zone of interest and returns the response (triggering the alert or not).

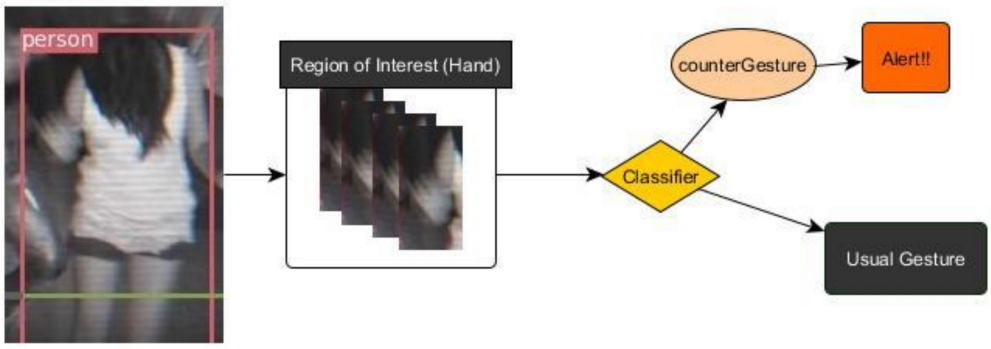

Figure 8. Considering the simplified chart, from left to right, we see how our model extracts the region of interest, processes the cropped images and classifies the gesture using the classifier. Nevertheless, the counterGesture integrated into the classifier block will trigger the alert if the specific gesture is identified three times.

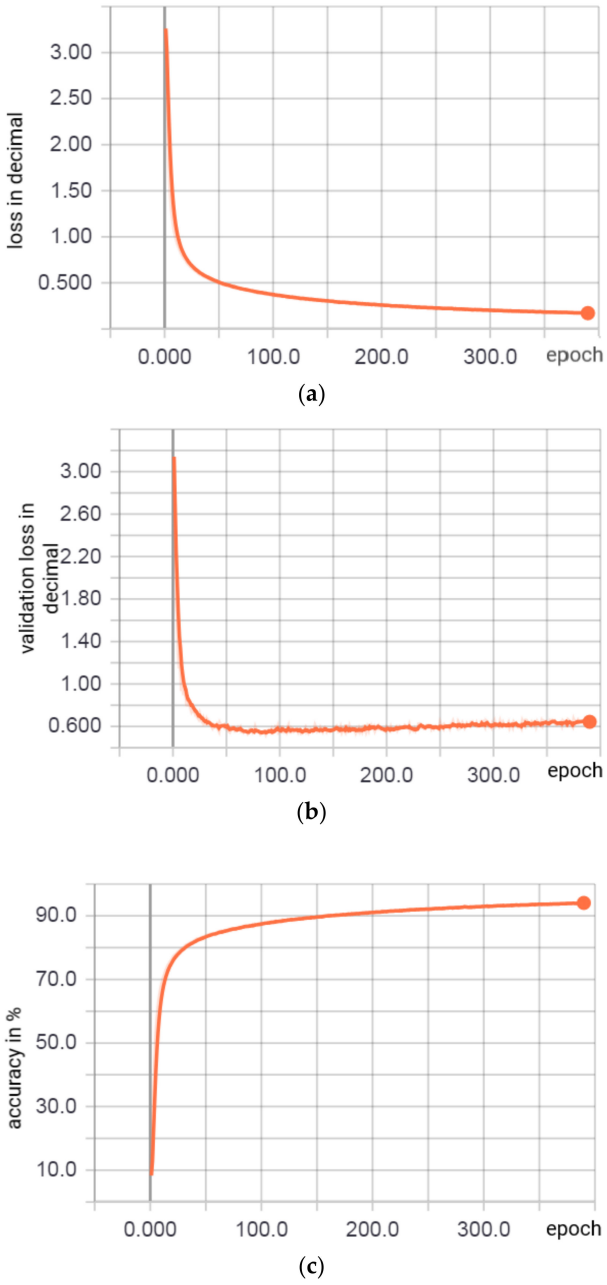

Figure 9. Performance of the classifier, where (**a**) represent the loss graph, where our classifier shows a decrease in the loss during the training phase; (**b**) represent the validation graph, which shows the reduction of the loss and the increase in the confidence of our classifier during the validation phase; and (**c**) show the accuracy graph, which shows an increase in accuracy of up to 91% after 350 epochs.

5. Results and Discussion

During the experiment, we firstly studied the performance of various versions of the neural network VGG-16 and a custom 3D CNN architecture on the classification task. In addition to this, we paid attention to the performance result of the number of input frames submitted to our gesture classification. Figure 9 provides an overview of how well our classifier performs, and the results in Table 4 show that we achieve a better performance by increasing the size of entries for all modalities. This depends strongly on the characteristics of the datasets used, especially the average duration of the gestures.

Table 4. Improvement of the custom 3D CNN by increasing the input size and comparison to state-of-the-art techniques applied on the EgoGesture dataset [45].

Model	Number of Input Frames	% Accuracy on RGB	% Accuracy on Depth
VGG-16 [47]	16	62.5	62.30
VGG-16 + LSTM [47]	16	74.60	77.69
Custom 3D CNN	16	85.1	87.9
Custom 3D CNN	24	90.3	90.6
Custom 3D CNN	32	91	91.3

Considering the video used in our experiment, let us have a close look at the portion of the video where the sexual assault is happening. Table 5 shows how Algorithm 1 (human detector) is applied on the video to detect a human, and it also gives an idea of when Algorithm 2 (hand gesture) detects the gesture. The hand gesture will activate the CounterGesture every time a specific gesture is identified. However, it is worth mentioning that in the case of a total occlusion of the hands of the victim, it is not possible for the system to detect the gesture and hence trigger the alert. As future work, we will explore the possibility of adding other factors to the decision making. Also, as shown in Table 5, the alarm is triggered after the counterGesture = 3.

Table 5. Step-by-step application of our algorithms on a sexual assault video.

Frames	Algorithm 1	Algorithm 2	CounterGesture	Alarm
	Human detected	Specific gesture detected	+1	Alarm triggered as CounterGesture= 3
	Human detected	Specific gesture detected	+1	
	Human detected	NA	0	
	Human detected	Specific gesture detected	+1	

Table 5. *Cont.*

Frames	Algorithm 1	Algorithm 2	CounterGesture	Alarm
	Human detected	Normal gesture detected	0	NA
	Human detected	NA	0	NA

When processing images from video surveillance, material quality (poor quality of image data, low light level, blur, pixelation of small objects) can be an obstacle; to address it, we propose enhancing the images (Figure 10) before submitting them to the human detector module (Figure 5).

Figure 10. Preprocessing of the images to ensure a better quality of input images in our system. (**a**) The original image, (**b**) the image on which we apply a normalization function to distribute the pixel intensities from 0 to 255; this will enhance the image without losing information.

Besides, the RGB-D (Depth sensor) frames are examined for different input sizes. Over the course of the experiment, the depth modality has proven to be essential to increase the performance, rather than a simple RGB input. Indeed, the depth component allowed the filtering of the movement from the background and helped to focus more on the movement of the hand, resulting in the discrimination of the features with the depth modality.

6. Conclusions

Considering sexual assault, this paper proposes a solution via the use of a new hierarchical architecture with three models for hand gesture alert systems. The proposed architecture enables efficient resource utilization and early detection for essential hand gesture alert applications. We obtained approximate results for both datasets when we evaluated our proposed model.

We defined a set of hands gestures that were identified by our classifier, and we introduced a module called "counterGesture". The latter allowed us to count the number of occurrences of a predefined gesture and trigger the alert. Besides, we found that the training time was far too long on the Jest dataset at a learning rate varying between 0.0001 and 0.001. We anticipate that in our future work, we will associate the facial expression with the alert decision in order to investigate ResNext [48–51] or a faster impact of the CNN-Region mode, at the same time as the detector and classifier [52,53]. Further study should consider the combination of lighter Deep neural networks (DNNs) to maintain accuracy and improve speed.

Author Contributions: Conceptualization, S.M., O.K., A.S. and K.K.; methodology, O.K., P.M. and K.K.; software, S.M. and O.K.; validation, O.K., S.M. and P.M.; formal analysis, A.S., S.M.; investigation, S.M.; resources, O.K., K.K. and P.M.; data curation, S.M., A.S. and O.K.; writing—original draft preparation, S.M. and O.K.; writing—review and editing, S.M., O.K., and A.S.; visualization, S.M., O.K..; supervision, A.S., K.K., P.M. and O.K.; project administration, O.K., K.K. and P.M.; funding acquisition, O.K. and K.K.

Funding: The work and the contribution were supported by project of excellence 2019/2205, Faculty of Informatics and Management, University of Hradec Kralove. The work was partially funded by the: (1) SPEV project, University of Hradec Kralove, FIM, Czech Republic (ID: 2103–2019), "Smart Solutions in Ubiquitous Computing Environments", (2) the Ministry of Education, Youth and Sports of Czech Republic (project ERDF no. CZ.02.1.01/0.0/0.0/18_069/0010054), (3) Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-20H04, Malaysia Research University Network (MRUN) Vot 4L876 and (4) the Fundamental Research Grant Scheme (FRGS) Vot 5F073 supported under Ministry of Education Malaysia for the completion of the research.

Acknowledgments: We are grateful for the support of student Sebastien Mambou and Michal Dobrovolny in consultations regarding application aspects. The APC was funded by project of excellence 2019/2205, Faculty of Informatics and Management, University of Hradec Kralove.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hust, S.J.T.; Rodgers, K.B.; Ebreo, S.; Stefani, W. Rape myth acceptance, efficacy, and heterosexual scripts in men's magazines: Factors associated with intentions to sexually coerce or intervene. *J. Interpers. Violence* **2019**, *34*, 1703–1733. [CrossRef] [PubMed]
2. Smith, S.G.; Chen, J.; Basile, K.C.; Gilbert, L.K.; Merrick, M.T.; Patel, N.; Walling, M.; Jain, A. *The National Intimate Partner and Sexual Violence Survey (NISVS): 2010-2012 State Report*; National Center for Injury Prevention and Control, Centers for Disease Control and Prevention: Atlanta, GA, USA, 2017.
3. Calafat, A.; Hughes, K.; Blay, N.; Bellis, M.A.; Mendes, F.; Juan, M.; Lazarov, P.; Cibin, B.; Duch, M.A. Sexual Harassment among Young Tourists Visiting Mediterranean Resorts. *Arch. Sex. Behav.* **2013**, *42*, 603–613. [CrossRef] [PubMed]
4. Haering, N.; Venetianer, P.L.; Lipton, A. The evolution of video surveillance: An overview. *Mach. Vis. Appl.* **2008**, *19*, 279–290. [CrossRef]
5. Elhamod, M.; Levine, M.D. Automated Real-Time Detection of Potentially Suspicious Behavior in Public Transport Areas. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 688–699. [CrossRef]
6. Nguyen, H.T.; Jung, S.W.; Won, C.S. Order-Preserving Condensation of Moving Objects in Surveillance Videos. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2408–2418. [CrossRef]
7. Hanson, G.C.; Perrin, N.; Moss, H.; Laharnar, N.; Glass, N. Workplace violence against homecare workers and its relationship with workers health outcomes: A cross-sectional study. *BMC Public Heal.* **2015**, *15*, 441.
8. Lätsch, D.C.; Nett, J.C.; Hümbelin, O. Poly-victimization and its relationship with emotional and social adjustment in adolescence: Evidence from a national survey in Switzerland. *Psychol. Violence* **2017**, *7*, 1–11. [CrossRef]

9. Yang, C.; Han, D.K.; Ko, H. Continuous hand gesture recognition based on trajectory shape information. *Pattern Recognit. Lett.* **2017**, *99*, 39–47. [CrossRef]
10. Traore, B.B.; Kamsu-Foguem, B.; Tangara, F. Deep convolution neural network for image recognition. *Ecol. Inform.* **2018**, *48*, 257–268. [CrossRef]
11. Pourbabaee, B.; Roshtkhari, M.J.; Khorasani, K. Deep Convolutional Neural Networks and Learning ECG Features for Screening Paroxysmal Atrial Fibrillation Patients. *IEEE Trans. Syst. Man Cybern. Syst.* **2018**, *48*, 2095–2104. [CrossRef]
12. Maron, H.; Galun, M.; Aigerman, N.; Trope, M.; Dym, N.; Yumer, E.; Kim, V.G.; Lipman, Y. Convolutional neural networks on surfaces via seamless toric covers. *ACM Trans. Graph.* **2017**, *36*, 1–10. [CrossRef]
13. Mambou, S.J.; Maresova, P.; Krejcar, O.; Selamat, A.; Kuca, K. Breast Cancer Detection Using Infrared Thermal Imaging and a Deep Learning Model. *Sensors* **2018**, *18*, 2799. [CrossRef] [PubMed]
14. Mambou, S.; Maresova, P.; Krejcar, O.; Selamat, A.; Kuca, K. Breast Cancer Detection Using Modern Visual IT Techniques. In *Modern Approaches for Intelligent Information and Database Systems*; Sieminski, A., Kozierkiewicz, A., Nunez, M., Ha, Q.T., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 397–407. ISBN 978-3-319-76081-0.
15. Mambou, S.; Krejcar, O.; Maresova, P.; Selamat, A.; Kuca, K. Novel Four Stages Classification of Breast Cancer Using Infrared Thermal Imaging and a Deep Learning Model. In *Bioinformatics and Biomedical Engineering*; Rojas, I., Valenzuela, O., Rojas, F., Ortuño, F., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 397–407. ISBN 978-3-319-76081-0.
16. Mambou, S.; Krejcar, O.; Selamat, A. Approximate Outputs of Accelerated Turing Machines Closest to Their Halting Point. In *Intelligent Information and Database Systems*; Nguyen, N.T., Gaol, F.L., Hong, T.P., Trawiński, B., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 702–713.
17. Mambou, S.; Krejcar, O.; Kuca, K.; Selamat, A. Novel Cross-View Human Action Model Recognition Based on the Powerful View-Invariant Features Technique. *Future Internet* **2018**, *10*, 89. [CrossRef]
18. Mambou, S.; Krejcar, O.; Kuca, K.; Selamat, A. Novel Human Action Recognition in RGB-D Videos Based on Powerful View Invariant Features Technique. In *Modern Approaches for Intelligent Information and Database Systems*; Sieminski, A., Kozierkiewicz, A., Nunez, M., Ha, Q.T., Eds.; Studies in Computational Intelligence; Springer International Publishing: Cham, Switzerland, 2018; pp. 343–353. ISBN 978-3-319-76081-0.
19. Ma, Y.; Deng, L.; Chen, X.; Guo, N. Integrating Orientation Cue With EOH-OLBP-Based Multilevel Features for Human Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 1755–1766. [CrossRef]
20. Tong, R.; Xie, D.; Tang, M. Upper Body Human Detection and Segmentation in Low Contrast Video. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 1502–1509. [CrossRef]
21. Jun, B.; Choi, I.; Kim, D. Local Transform Features and Hybridization for Accurate Face and Human Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1423–1436. [CrossRef]
22. 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; 2005; Volume 1, pp. 886–893.
23. Hedged Deep Tracking. Available online: https://www.computer.org/csdl/proceedings-article/cvpr/2016/07780835/12OmNrHjqOQ (accessed on 26 May 2019).
24. Wang, L.; Ouyang, W.; Wang, X.; Lu, H. Visual Tracking with Fully Convolutional Networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3119–3127.
25. Fu, Z.; Angelini, F.; Naqvi, S.M.; Chambers, J.A. GM-PHD Filter Based Online Multiple Human Tracking Using Deep Discriminative Correlation Matching. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4299–4303.
26. He, F.; Wu, Y.; Yi, S.; Wang, X.; Wang, H.; Liu, W.; Feng, B. Depth-Projection-Map-Based Bag of Contour Fragments for Robust Hand Gesture Recognition. *IEEE Trans. Hum. Mach. Syst.* **2017**, *47*, 511–523.
27. Mo, Z.; Neumann, U. Real-time hand pose recognition using low-resolution depth images. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1499–1505.
28. Ren, Z.; Yuan, J.; Zhang, Z. Robust Hand Gesture Recognition Based on Finger-earth Mover's Distance with a Commodity Depth Camera. In Proceedings of the 19th ACM International Conference on Multimedia, New York, NY, USA, 28 November–1 December 2011; pp. 1093–1096.

29. Liu, K.; Gong, D.; Meng, F.; Chen, H.; Wang, G.G. Gesture segmentation based on a two-phase estimation of distribution algorithm. *Inf. Sci.* **2017**, *394–395*, 88–105. [CrossRef]
30. Erol, A.; Bebis, G.; Nicolescu, M.; Boyle, R.D.; Twombly, X. Vision-based Hand Pose Estimation: A Review. *Comput. Vis. Image Underst.* **2007**, *108*, 52–73. [CrossRef]
31. Oikonomidis, I.; Kyriazis, N.; Argyros, A. Efficient model-based 3D tracking of hand articulations using Kinect. In *British Machine Vision Conference 2011*; British Machine Vision Association: Dundee, UK, 2011; pp. 101.1–101.11.
32. Tang, M. *Recognizing Hand Gestures with Microsoft's Kinect*; Stanford University: Stanford, CA, USA, 2011.
33. Bai, X.; Bai, S.; Zhu, Z.; Latecki, L.J. 3D Shape Matching via Two Layer Coding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1. [CrossRef]
34. Bai, X.; Latecki, L. Path Similarity Skeleton Graph Matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1282–1292.
35. De Smedt, Q.; Wannous, H.; Vandeborre, J.-P. Heterogeneous hand gesture recognition using 3D dynamic skeletal data. *Comput. Vis. Image Underst.* **2019**, *181*, 60–72. [CrossRef]
36. Yang, J.; Li, S.; Gao, Z.; Wang, Z.; Liu, W. Real-Time Recognition Method for 0.8 cm Darning Needles and KR22 Bearings Based on Convolution Neural Networks and Data Increase. *Appl. Sci.* **2018**, *8*, 1857. [CrossRef]
37. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 8 June 2016; pp. 779–788.
38. Al-masni, M.A.; Al-antari, M.A.; Park, J.M.; Gi, G.; Kim, T.Y.; Rivera, P.; Valarezo, E.; Choi, M.T.; Han, S.M.; Kim, T.S. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Comput. Methods Programs Biomed.* **2018**, *157*, 85–94. [CrossRef]
39. Tang, C.; Ling, Y.; Yang, X.; Jin, W.; Zheng, C. Multi-View Object Detection Based on Deep Learning. *Appl. Sci.* **2018**, *8*, 1423. [CrossRef]
40. Yang, G.; Yang, J.; Sheng, W.; Junior, F.E.F.; Li, S. Convolutional Neural Network-Based Embarrassing Situation Detection under Camera for Social Robot in Smart Homes. *Sensors* **2018**, *18*, 1530. [CrossRef]
41. Jiang, M.; Hai, T.; Pan, Z.; Wang, H.; Jia, Y.; Deng, C.; Yu, Y.; Shan, J. Multi-Agent Deep Reinforcement Learning for Multi-Object Tracker. *IEEE Access* **2019**, *7*, 32400–32407. [CrossRef]
42. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
43. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE international Conference on Computer Vision, Washington, DC, USA, 7–13 December 2014.
44. He, L.; Xu, L.; Ming, X.; Liu, Q. A Web Service System Supporting Three-dimensional Post-processing of Medical Images Based on WADO Protocol. *J. Med Syst.* **2015**, *39*, 39. [CrossRef]
45. The 20BN-JESTER Dataset—Twenty Billion Neurons. Available online: https://20bn.com/datasets/jester/v1#download (accessed on 26 May 2019).
46. Mumbai Girl Molesting CCTV footage—YouTube. Available online: https://www.youtube.com/watch?v=D8z1DAIo_Lg (accessed on 26 May 2019).
47. Zhang, Y.; Cao, C.; Cheng, J.; Lu, H. EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition. *IEEE Trans. Multimedia* **2018**, *20*, 1038–1050. [CrossRef]
48. Balagopal, A.; Kazemifar, S.; Nguyen, D.; Lin, M.-H.; Hannan, R.; Owrangi, A.M.; Jiang, S.B. Fully automated organ segmentation in male pelvic CT images. *Phys. Med. Boil.* **2018**, *63*, 245015. [CrossRef]
49. Wang, Y.; Wang, Z. A survey of recent work on fine-grained image classification techniques. *J. Vis. Commun. Image Represent.* **2019**, *59*, 210–214. [CrossRef]
50. Le, T.H.; Huang, S.C.; Jaw, D.W. Cross-Resolution Feature Fusion for Fast Hand Detection in Intelligent Homecare Systems. *IEEE Sens. J.* **2019**, *19*, 4696–4704. [CrossRef]
51. Zhaodi, W.; Menghan, H.; Guangtao, Z. Application of Deep Learning Architectures for Accurate and Rapid Detection of Internal Mechanical Damage of Blueberry Using Hyperspectral Transmittance Data. *Sensors* **2018**, *18*, 1126. [CrossRef]

52. Chen, Y.P.; Li, Y.; Wang, G.; Xu, Q. A Multi–Strategy Region Proposal Network. *Expert Syst. Appl.* **2018**, *113*, 1–17. [CrossRef]
53. Cheng, G.; Han, J.; Zhou, P.; Xu, D. Learning Rotation-Invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection. *IEEE Trans. Image Process.* **2019**, *28*, 265–278. [CrossRef]