Contagion Prevention of COVID-19 by means of Touch Detection for Retail Stores

Rafał Brociek[1], Giorgio De Magistris[2], Francesca Cardia[2], Federica Coppa[2] and Samuele Russo[3]

[1]Sapienza University of Rome, piazzale Aldo Moro 5, Roma 00185, Italy
[2]Department of Computer, Automation and Management Engineering, Sapienza University of Rome, via Ariosto 25 Roma 00185, Italy
[2]Department of Psychology, Sapienza University of Rome, via dei Marsi 78 Roma 00185, Italy

Abstract

The recent Covid-19 pandemic has changed many aspects of people's life. One of the principal preoccupations regards how easily the virus spreads through infected items. Of special concern are physical stores, where the same items can be touched by a lot of people throughout the day. In this paper a system to efficiently detect the human interaction with clothes in clothing stores is presented. The system recognizes the elements that have been touched, allowing a selective sanitization of potentially infected items. In this work two approaches are presented and compared: the pixel approach and the bounding box approach. The former has better detection performances while the latter is slightly more efficient.

1. Introduction

The recent Covid-19 pandemic has affected hardly most commercial activities[18, 5, 23]. The recent restrictions imposed by the governments to contrast the virus spreading had a big impact on most retail stores, favouring online shopping [14], where the infection risk through infected items is obviously reduced. In this context, an efficient sanitization of stores would decrease the exposure to infection [6, 2] making people more inclined to return to physical shopping.

Some contexts, especially where several people are present at the same time, often do not allow to keep under control every part of the environment. In particular it gets difficult to stay aware about all the physical contacts with people, among themselves, with the environment and with its content. During the COVID-19 pandemic scenario it has become necessary to constantly sanitize the environment and all its potentially contaminated parts. Therefore it has become clear that very help that can facilitate this task would be of great use in such contexts. Moreover the sanitizing actions carried out by an employee in the presence of the customer, in certain circumstances, can induce a feeling of annoyance or discomfort. However, postponing the intervention can be difficult because the cleaner would not remember precisely which parts of the environment came into contact with the customer and the same cannot take them all into consideration because at that moment he was not present or the his attention was focused elsewhere. The use of an automatic system capable to recognize and remember potentially contaminated areas or objects can considerably reduce the effort for the sanitizer and improve the accuracy and effectiveness of his action. At the same time, the implementation of such a solution would considerably reduce the feeling of discomfort that the customer can experience in the presence of a sanitizer who disinfects every object that the customer has touched in front of the customer. This aspect allows the customer avoid embarrassments while maintaining a relationship of trust, reducing the risks of a mortification, for which the customer would feel limited by the possibility of expressing his own behavior while exploring the store. This principle can also be applicable to professional studies or other facilities, where the construction of an alliance and a relationship of trust between the professional and the client is always a critical and delicate moment that must be managed with the utmost sensibility.

The aim of this work consists in creating a system that is able to help the sellers to sanitize items faster and more efficiently, knowing which product should be sanitized and which do not. In particular, we designed a system for clothing stores, but the same solution can be adapted to many other retail stores. The general idea consists in the implementation of a system that is able to detect the touch action. We decided to restrict the context to clothing stores because the model is more efficient when trained on a specific set of objects. Moreover, clothing stores represent one of the commercial activities with a higher risk of Covid bacteria spreading, since people touch and try on dresses continuously before buying them.

In section 2 we formalize the problem of *touch detection* and we relate it to the state of the art. Section 3 and 4

SYSTEM 2021 @ Scholar's Yearly Symposium of Technology, Engineering and Mathematics. July 27–29, 2021, Catania, IT
rafal.brociek@polsl.pl (R. Brociek);
demagistris@diag.uniroma1.it (G. De Magistris);
cardia.1759331@studenti.uniroma1.it (F. Cardia);
coppa.1749614@studenti.uniroma1.it (F. Coppa);
samuele.russo@uniroma1.it (S. Russo)
0000-0002-7255-6951 (R. Brociek); 0000-0002-3076-4509 (G. De Magistris); 0000-0002-1846-9996 (S. Russo)

respectively describe the models and the datasets that are used in the proposed system. In section 5 we illustrate the training strategy and some implementation details in order to make the system easily reproducible. In section 6 we report the performance metrics and compare the different approaches. In section 7 conclusions are drawn.

2. Problem Definition and State of Art

This paper presents a new method for detecting the "touch" event and in particular we narrowed the scope to the action of touching clothes with one's hands. The collision detection task in a 3D environment is a well studied problem [16, 19, 15] in literature and it finds application in many fields such as robotics and video gaming. However, to the best of our knowledge, there is no equivalent formulation in the context of 2D images, where there is no depth information. According to our formulation, the touch detection is based on the recognition of the objects of interest (in this case clothes and hands). The result of such recognition, depending on the method that has been used, can be either a set of coordinates identifying a bounding box (detection) or a pixel mask (segmentation). The bounding box or the pixel mask is then used to check if there is an overlap between the two objects. We will refer to the former as *Bounding Box approach* and to the latter as *Pixel approach*. In the first case a simple check on the coordinates is sufficient (see algorithm 1), while in the other a parallel scan of the pixels of the two masks is needed (see algorithm 2).

Algorithm 1: Algorithm to check the overlap between two rectangular bounding boxes

Input:
A,B = upper-left and bottom-right vertices of the first rectangle
A',B' = upper-left and bottom-right vertices of the second rectangle
Output:
Overlap / No Overlap
Time Complexity:
$\mathcal{O}(1)$
Algorithm
if *A'.x > B.x or A.x > B'.x* **then**
  return *No Overlap*
if *B.y > A'.y or B'.y > A.y* **then**
  return *No Overlap*
return *Overlap*

Algorithm 2: Algorithm to check the overlap between two pixel masks

Input:
$M1_{N\times N}$ = pixel mask of the first object
$M2_{N\times N}$ = pixel mask of the second object
Output:
Overlap / No Overlap
Time Complexity:
$\mathcal{O}(N^2)$
Algorithm
for *i = 0 to N-1* **do**
  for *j = 0 to N-1* **do**
    if $M1_{i,j}$ *and* $M2_{i,j}$ **then**
      return *Overlap*
return *No Overlap*

3. Method

Object detection and image segmentation are two foundamental problems in computer vision. Before the incredible success of deep learning these tasks were performed using solely standard computer vision algorithms. For example the selective search [24, 20] leverages the hierarchical structure of images and, from an initial segmentation, it recursively merges similar patches in terms of color, texture, size and shape[**Capizzi201645**, 4]. State of the art deep learning models for detection and segmentation are based on the R-CNN architecture introduced in [10]. This network receives as input a set of region proposals which are the candidates for the classifications, the architecture is independent of the algorithm used, then a pre-trained large CNN network is used to extract features from the selected regions and then class specific Linear Support Vector Machines (SVM) are used to classify the regions[21]. The main problem of this architecture was the long evaluation time, preventing the model from online usage, hence Fast R-CNN [9] was introduced to speed-up evaluation time. This model learns to classify object proposals and to refine their spatial locations jointly. Each region proposal is mapped into a fixed-length feature vector using interleaved convolutional and pooling layers followed by fully connected layers. Then the feature vector flows into the two output branches which outputs are respectively: softmax probabilities and per-class bounding-box regression offsets. The architecture is trained end-to-end with a multi-task loss.

For the task of object detection, we used Faster R-CNN [22] that is an extension of Fast R-CNN that avoids the bottleneck of the region proposal module with the introduction of a Region Proposal Network (RPN). The RPN is a fully convolutional network sharing the convolutional features of the detection network that simultaneously

predicts object bounds and objectness scores at each position. It is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection.

For the task of image segmentation, we used Mask R-CNN [13] that extends the Faster R-CNN architecture with a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. This network adds only a small overhead with respect to Faster R-CNN and it runs at 5 fps. Moreover Mask R-CNN surpasses all previous state-of-the-art single-model results on the COCO instance segmentation competition [17].

4. Datasets

The models from the R-CNN family are trained with labelled and annotated images. We trained two separate models, respectively for hands and clothes recognition, hence the objects of interest are labelled with a single label. For the task of object detection the annotation consists of the four values that identify the bounding box that are the x and y pixel coordinates of the center, width and height in pixels. For the segmentation task the ground truth is another image with the same dimensions of the original image where the pixels that belongs to the object of interest are white (mask) and the background is black (see figure 1). The network only allows dimensions

Figure 1: Pixel masks for a hand (left) and for trousers

like 256,320,384 or whatever is dividable by 2 at least 6 times. For this reason each image in both hands and Clothing datasets have dimensions 384×448. The Hands Dataset (see figure 2) is obtained collecting 400 images for training and 100 for testing from three famous datasets: EgoHands [3], HandOverFace [8] and EgoYouTubeHands [25]. Moreover 40 images for training and 10 for testing were added manually. We chose images from multiple datasets to have representations of hands in different contexts, in order to improve the generalization power of the model. For the clothing recognition task we built a dataset of 500 images labelled with the following four labels (see figure 3): t-shirt, trousers, skirt, long sleeve. We followed the common practice of partitioning the dataset using the 80% for training and the remaining

Figure 2: Samples from the Hands Dataset

for validation. Some images were randomly selected trough Google Search, some were taken from the known Clothing Dataset [11] and others were added manually. In both datasets, images were annotated using the VIA Annotation Software [7] that is an open source light weight software that runs in the web browser and allows to annotate images with bounding boxes or pixel masks.

Figure 3: Samples from each of the four categories of the Clothing dataset, from left to right: t-shirt, trousers, skirt, long sleeve

5. Training

We trained the two models separately, respectively for hands and clothes detection and segmentation. We used the Mask R-CNN implementation provided by [1] both for detection (bounding box) and segmentation. Remember that Mask R-CNN is an extension of Faster R-CNN that adds a branch for predicting the mask, but the rest of the architecture is unchanged, including the branch for the bounding box regression. This implementation requires only the annotation with the pixel mask, the bounding box for the ground truth is computed on the fly picking the smallest box that encapsulates all the pixels of the mask. Both models have been fine tuned (all layers) for 50 epochs, with a learning rate of 0.0001, a weight decay of 0.00001, ResNet-101 [12] as Backbone and some data

augmentation techniques to improve the performances of Mask R-CNN. Figure 4 shows the learning curves, while in figure 5 the single components of the loss function are illustrated for the validation set. Considering that the two models have similar plots, we will illustrate only those regarding the model trained on the Clothing dataset for conciseness sake.

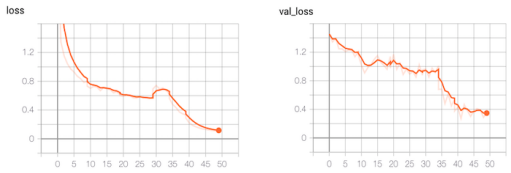

Figure 4: Learning curves, from left to right the training and validation loss

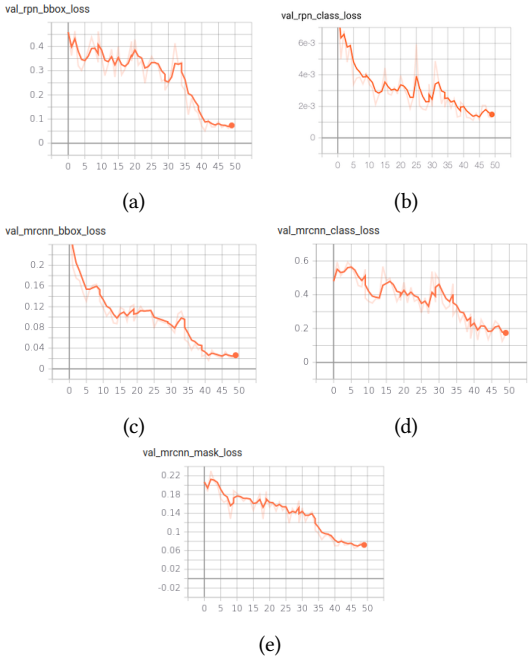

Figure 5: Components of the loss function of the Mask R-CNN. The rpn_class_loss 5a and rpn_bbox_loss 5b indicates respectively how well the Region Proposal Network separates background from objects and localizes objects. While the mrcnn_bbox_loss 5c, mrcnn_class_loss 5d and mrcnn_mask_loss 5e measure the performances of the Mask R-CNN in localizing, labelling and segmenting objects.

6. Results

At this point we have the two models trained to detect with a bounding box and segment respectively clothes and hands. In order to test the two approaches (bounding box vs pixel mask) we built manually a new dataset with 100 photographed images with hands and clothes and we labelled each image with the two labels *overlap* and *no overlap*. In order to check the overlap we used algorithms 1 and 2 respectively for the bounding box and pixel approaches. The result is a set of images with their associated labels. Table 1 reports some metrics that are commonly used to evaluate the detection, while figure 6 shows the confusion matrices for the two approaches.

	Accuracy	Precision	Recall	F1 score
Bounding Box	0.56	0.53	0.89	0.66
Pixel Mask	0.88	0.89	0.88	0.88

Table 1
Evaluation metrics for the task of touch detection. In the first row the metrics for the Bounding Box approach while in the second row the ones for the Pixel approach

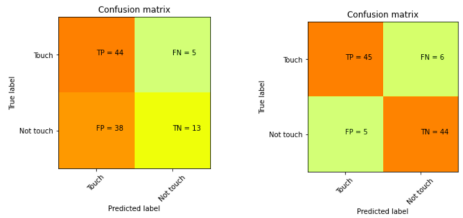

Figure 6: The confusion matrices for the touch detection task. On the left the values refer to the Bounding Box approach, on the right they refer to the Pixel approach.

From these metrics it emerges that the Pixel approach is much superior than the Bounding Box approach. In particular, the Bounding Box approach returns a lot of false positives, because often the bounding boxes overlap while the objects inside do not, as shown in figure 7.

Figure 7: Bounding box approach (left) compared with the Pixel approach (right). This is an example misclassification by the Bounding Box approach (the rectangles are contained one into the other) and correct classification by the Pixel based approach (the pixel masks have no pixel in common).

7. Conclusion

In this work a system to efficiently detect the human interaction with objects in clothing stores was presented. The proposed system can be easily adapted to a variety of other fields by changing the datasets used for the object detection and segmentation tasks. We presented two approaches, the former based on object detection with bounding box and the latter based on segmentation, and we showed that the second one performed much better with the cost of a small overhead. A further improvement to the proposed model would be the introduction of depth information. This extension however, would increase the performances at the expense of a higher cost for more specialized hardware and this factor could limit its widespread use. That said, we think that our system achieves good enough results to be implemented in physical stores as a highly cost-effective tool for the Covid-19 pandemic containment.

References

[1] Waleed Abdulla. *Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow*. https://github.com/matterport/Mask_RCNN. 2017.

[2] R. Avanzato et al. "YOLOv3-based mask and face recognition algorithm for individual protection applications". In: vol. 2768. 2020, pp. 41–45.

[3] Sven Bambach et al. "Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions". In: *The IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.

[4] F. Bonanno et al. "Optimal thicknesses determination in a multilayer structure to improve the SPP efficiency for photovoltaic devices by an hybrid FEM - Cascade Neural Network based approach". In: 2014, pp. 355–362. DOI: 10.1109/SPEEDAM.2014.6872103.

[5] P. Caponnetto et al. "The effects of physical exercise on mental health: From cognitive improvements to risk of addiction". In: *International Journal of Environmental Research and Public Health* 18.24 (2021). DOI: 10.3390/ijerph182413384.

[6] Federica Carraturo et al. "Persistence of SARS-CoV-2 in the environment and COVID-19 transmission risk from environmental matrices and surfaces". In: *Environmental pollution* 265 (2020), p. 115010.

[7] Abhishek Dutta and Andrew Zisserman. "The VIA annotation software for images, audio and video". In: *Proceedings of the 27th ACM international conference on multimedia*. 2019, pp. 2276–2279.

[8] Sakher Ghanem, Ashiq Imran, and Vassilis Athitsos. "Analysis of hand segmentation on challenging hand over face scenario". In: *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 2019, pp. 236–242.

[9] Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[10] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[11] Alexey Grigorev. *Clothing dataset*. 2020. URL: https://www.kaggle.com/agrigorev/clothing-dataset-full.

[12] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[13] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[14] Rae Yule Kim. "The impact of COVID-19 on consumers: Preparing for digital sales". In: *IEEE Engineering Management Review* 48.3 (2020), pp. 212–218.

[15] Sinan Kockara et al. "Collision detection: A survey". In: *2007 IEEE International Conference on Systems, Man and Cybernetics*. IEEE. 2007, pp. 4046–4051.

[16] Ming Lin and Stefan Gottschalk. "Collision detection between geometric models: A survey". In: *Proc. of IMA conference on mathematics of surfaces*. Vol. 1. Citeseer. 1998, pp. 602–608.

[17] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[18] V. Marcotrigiano et al. "An integrated control plan in primary schools: Results of a field investigation on nutritional and hygienic features in the apulia region (southern italy)". In: *Nutrients* 13.9 (2021). DOI: 10.3390/nu13093006.

[19] C. Napoli, G. Pappalardo, and E. Tramontana. "An agent-driven semantical identifier using radial basis neural networks and reinforcement learning". In: vol. 1260. 2014.

[20] C. Napoli, G. Pappalardo, and E. Tramontana. "Using modularity metrics to assist move method refactoring of large systems". In: 2013, pp. 529–534. DOI: 10.1109/CISIS.2013.96.

[21] B.A. Nowak et al. "Multi-class nearest neighbour classifier for incomplete data handling". In: vol. 9119. 2015, pp. 469–480. DOI: 10.1007/978-3-319-19324-3_42.

[22] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015), pp. 91–99.

[23] S. Russo et al. "Reducing the psychological burden of isolated oncological patients by means of decision trees". In: vol. 2768. 2020, pp. 46–53.

[24] Jasper RR Uijlings et al. "Selective search for object recognition". In: *International journal of computer vision* 104.2 (2013), pp. 154–171.

[25] Aisha Urooj and Ali Borji. "Analysis of hand segmentation in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4710–4719.