Computer vision with cognitive learning to improve the decision-making during the sales process in physical stores

Vinícius da Silva Ramalho[1], Anderson Luis Szejka[1][0000-0001-8977-1351], Marcelo Rudek[1] [0000-0002-6170-3370] and Osiris Canciglieri Junior[1][0000-0002-8503-9275]

[1] Industrial and Systems Engineering Graduate Program, Pontifical Catholic University of Parana, Curitiba, Brazil
{vinicius.ramalho, anderson.szejka, marcelo.rudek, osiris.canciglieri}@pucpr.br

Abstract. In a world where the information is obtained faster than ever seem, new methods to process that high volume of data are being developed frequently. This is more notorious in a virtual ambient where the data is generated in a manner that is faster and easier to analyze than in the real world. This is very evident in the retail field, where virtual stores have easy access to all the advertisement a user visited and simple to obtain user profile, on the other hand physical stores are limited to basically create a register in a database when there is a purchase. In an attempt to improve the retailers experience from physical stores to manage their business this document has the objective to develop a computational tool that will analyze the people flux going in the establishment, trying to inform the retailer the amount of people and their gender to help the sales process in physical stores. To this end, computational vision methods and algorithms were raised, which after selection, theoretical conception and tool's implementation it was tested with benchmarks to operate locally and in real time by accessing the cameras installed strategically in a real scenario. Two scenarios were tested: static ambient light and dynamic light. Two tests were conducted: YOLOv2 against background subtraction-based counter; gender classification using full body features. Even though the results were not as positive as needed for commercial use, the tool demonstrated potential and space for improvements.

Keywords: Smart retail. Retail 4.0. Computer Vision. Intelligent systems.

1 Introduction

The development of a retail store is directly linked to the customers' experience, interests and exclusiveness therefore gathering more information about customers will help the retailer give the consumer a better experience and improve business growth [1]. To provide such an experience a valuable insight is needed to better understand client's shopping profile [2].

Consequently, physical stores are in a disadvantage when compared with virtual stores as the later can better understand their customers. Because the data is already in

the digital world, virtual stores can easily access the search history, gender, age, nationality and more, which are very important to figure out client's profiles [3].

Using new tools like intelligent systems, computer vision, and internet of things (IoT), the retail 4.0 emerges as a way to lessen the difference between buying online and offline. The retail 4.0 brings more attention to digital marketing and social media, dynamic pricing of the products accordingly to demand and stock volume and, by understanding the customers profile, provide a customized experience. It is in the interest of physical stores to make a smart marketing campaign using client's data to optimize the selling process [4].

Based on this context, this paper proposes a computational tool utilizing computer vision technics to extract and provide the number of clients and their gender in real time as well as test the tool in real scenarios with benchmarks to better analyze the approach.

The tool will provide raw data, so it will need to be refined or crossed with other data. The number of clients can be cross related with weather or holidays to forecast how many clients is to be expected, so the owner can better adjust employees scheduling and stock management. Gender will give a better hint on customers taste to help, for instance, the selection of a promotional product. While this work focus on those two parameters age and emotions are also possible as demonstrated in [5].

2 Technical background

The object detection technics can be separated in 4 classes: feature-based; template-based; movement-based; classifier-based. For its simplicity and capacity to process in real time the movement-based approach known as background subtraction was selected as well as a more robust, however computing intensive, convolutional neural network approach.

2.1 Background subtraction

Easily found in safety and security systems to detect moving objects. A scene model is built pixel by pixel to verify, using pixel value comparison, differences in the model and the actual frame. If the difference Is very small then it is probably just noise in the image, however if it is large that means it is a moving object [6].

The background model could be a static image (when the scene is empty), a pixel wise mean/median of passed frames, a Gaussian distribution approach and many more. After the subtraction it is applied a threshold to classify whether it is background or foreground to obtain the final result. The process described can be easily expressed in the equation 1 [7].

$$result(x,y) = \begin{cases} 1, |I(x,y) - model(x,y)| > threshold \\ 0, |I(x,y) - model(x,y)| \leq threshold \end{cases} \quad (1)$$

Where $I(x,y)$ is the actual frame, $model(x,y)$ is the background's model and $result(x,y)$ is a bitmap that describes if a pixel belongs to the foreground (value "1") or background (value "0").

The method to model the background in this paper is based on the mean value, pixel wise, of n passed frames which is a quick technic but usually uses more memory than other methods [8]. This method is indicated for situations where the camera changes position. The background's model is obtained using the equation 2.

$$model(x,y,z) = \frac{1}{n}\sum_i^n history(i,x,y,z) \quad (2)$$

Where $history$ is a 4-dimension hyper volume, that for each position i there is a volume representing an image with width x, height y and z color channels (usually 3) as represented bellow.

$$history(i,x,y,z) = [I_0(x,y,z), I_1(x,y,z) \dots, I_i(x,y,z) \dots I_n(x,y,z)] \quad (3)$$

2.2 YOLOv2

The computer vision improved a lot after applications using convolutional neural networks (CNN) for object detection were developed. Using the brain as an inspiration a CNN can obtain great performance on pattern recognition tasks using extracted features learned beforehand. The difference between a CNN and a conventional artificial neural network (ANN) is that neurons are organized in three dimensions, where the height and width receive input signals and depth is an activation volume. Any neuron connects to a small region on the following layer which means that the image information will be condensed to a smaller volume [9].

There are plenty CNN architectures, but when it comes to real-time applications only some architectures can be realistically used. Using a clever approach to object detection YOLO is a very fast CNN architecture, unlike other methods like R-CNN that uses, in different steps, a feature extractor, classifier and a regressor to both detect and classify object in an image [10]. Instead, YOLO approach the whole problems as a regression problem from the image's pixel to the bounding boxes coordinates which makes it faster and easier to train [11][12].

3 Computational tool's architecture

The computational tool proposed in this paper is software that will operate in a local machine processing videos from strategically placed cameras in real-time to count and classify the customers gender.

3.1 Camera's layout

The software should be able to analyze the flux of people walking through the door with the least amount of occlusion as possible. To that end one camera should be positioned above the door and capturing the ground and people walking through. As the

characteristics captured from above are not enough to classify gender another camera in front of the door, but away from the path people use, is needed to capture more features and closer.

3.2 General working

The software is divided in two modules that work in parallel: count module, and classification module. In order to minimize the computational cost, the classification module will only engage when the count module detects someone going inside the room, when that happens a sequence of frames will be saved to be processed by the classification module. That will prevent the neural network from working on every single frame, but only on those that really matters.

The count module will have 2 options to detect people: the background subtraction and the YOLOv2-based CNN. The idea is to be able to work on modest machines as well as on more powerful computers, and as the count module must work on every single frame using the resources most the time it is the only module to receive those options. The tracking is based on the intersection of the bounding boxes along adjacent frames, the pairs with most intersected area are considered the same person throughout the frame sequence. The counting is based on a single line in the middle of the frame, if the center of the bounding box was above the line and in the next frame it appears bellow the line it is counted down. If the order is inverted the bounding box is counted up, in that case, it also saves the frames from the other camera in front of the door to be classified by the classification module.

The classification module will basically stand by until frames are saved to be processed. The module will run a different YOLOv2-based CNN trained to detect people and classify their gender using information of the whole body, instead of using face features like other methods. Face is not that easy to capture in many situations, people might look away, look at their phones, put their hands on their faces and many more, so to be able to classify in a more robust way the network was trained with the whole body. To improve the robustness, the CNN will classify a sequence of frames and if the person is classified more times as a male then the module will consider a male and vice versa.

4 Training

The first YOLOv2-based CNN was modified to detect one class (person) and the second CNN two classes (male and female), that was done by changing the number of filters in the last layer in each CNN to 30 and 35 respectively. For people detection the CNN was trained on COCO dataset plus 2800 custom images, the custom images were from cameras positioned above doors from 2 different places. It was hoped that those examples would improve detections from above the door.

For male and female detections using the full body features a dataset was created based on the WIKI and IMDB datasets. The WIKI and IMDB datasets only have the face coordinates annotated, and using the other CNN trained for people detection face

coordinates were updated for the whole body. The problem is that there is always one person annotated per image in the dataset, as it is concentrated on specific personnel, causing some images to have people missing annotations. Those images were used for training ONLY. It was also added customized PASCAL VOC dataset with manually labeled gender in each image, and a chunk of those images were the only ones used to evaluate the CNN (achieved a classification F1score of 0.70).

5 Benchmarks explanation

To measure the computational tool's performance two benchmarks were created to test each module. The benchmarks use manually annotated bounding boxes to compare with the ones the algorithm is outputting. The metrics selected were precision, recall and F1score which are vastly used to measure classification performance. A high precision means that the model is good at classifying examples of that class and recall measures how much the actual class (all data) was classified correctly. Those 2 metrics are joined to obtain F1score. The principal advantage of F1score is the ability to measure unbalanced data which means, for instance, the number of false positives doesn't need to be the same or close to false positives [13] making it easier to evaluate real scenarios, where there are more variables, balanced data is hardly acquired.

6 Real world tests

To prove and test the proposed tool, it will be tested in 2 different scenarios. The first scenario, university laboratory, the light will be static and the second scenario, a real store, the light will be dynamic and unpredictable. The first test will compare traditional and cognitive computer vision to verify if the YOLO-based counter is better than the background subtraction approach and how much is the difference. The last will analyze the gender counting performance. 4 videos, with 1 hour each, in 4 different cameras were recorded, 2 for each scenario, and manual annotations were made frame by frame in all 16 hours of video.

6.1 Comparison between traditional and cognitive computer vision

The scenario with controlled light the line was positioned in the middle of the frame close to the door to guarantee good visibility since before the person walking in. Calculating the selected performance metrics for each video the Table 1 is built with all results and combined, each direction has 3 metrics per video and a combined result of all the videos in the last row.

Table 1. Counting benchmark's metrics (traditional approach) in the controlled light scenario

Direction	Up			Down		
Metrics	Precision	Recall	F1score	Precision	Recall	F1score
Video 1	0.9999	0.9333	0.9655	0.8333	0.9090	0.8695

Video 2	0.3181	0.6363	0.4224	0.2258	0.5833	0.3255
Video 3	0.9999	0.7499	0.8571	0.7999	0.8888	0.8421
Video 4	0.9999	0.9999	0.9999	0.7499	0.8571	0.7999
Comb.	0.6739	0.8157	0.7380	0.5081	0.7948	0.6199

The F1score decay a lot in the second video making the combined result decrease. This bad performance issue was cause because the door was opened and closed, Fig. 1, a lot throughout the second video. This situation creates differences in the modeled background causing false positives.

Fig. 1. Opening and closing door example causing false positives

Running the benchmarks using YOLO as detector the Table 2 was built. Even though the F1score is a little bit higher than the traditional approach the recall for people going out was low because the detector failed to detect people. In contrast the F1score increased more than 6% for people going up (inside), this improvement is visible in the second video where the door is not counted because YOLO know it is not a person resulting in better precision.

Table 2. Counting benchmark's metrics (cognitive approach) in the controlled light scenario

Direction	Up			Down		
Metrics	Precision	Recall	F1score	Precision	Recall	F1score
Video 1	0.9166	0.7333	0.8148	0.8888	0.7272	0.7999
Video 2	0.9999	0.7272	0.8421	0.8749	0.5833	0.6999
Video 3	0.8333	0.6249	0.7142	0.6666	0.2499	0.3636
Video 4	0.9999	0.7499	0.8571	0.9999	0.2857	0.4444
Comb.	0.9310	0.71	0.8059	0.8336	0.4999	0.6333

In the dynamic light scenario, the camera was installed not as close to the door causing the line to be moved to 75% the frame's height. Running the benchmark in this scenario with the traditional approach the metrics, Table 3, were obtained.

Table 3. Counting benchmark's metrics (traditional approach) in the dynamic light scenario

Direction	Up			Down		
Metrics	Precision	Recall	F1score	Precision	Recall	F1score
Video 1	0.4074	0.8461	0.5499	0.4615	0.9230	0.6153
Video 2	0.7499	0.9999	0.8571	0.6666	0.9999	0.7999
Video 3	0.9999	0.8999	0.9473	0.9999	0.6666	0.7999
Video 4	0.7142	0.7142	0.7142	0.7499	0.7499	0.7499
Comb.	0.5957	0.8484	0.6999	0.6046	0.8124	0.6933

In those videos the door was not moved as much, but during the cleaning process the squeegee, Fig. 2, was counted several times causing a decreasing precision.

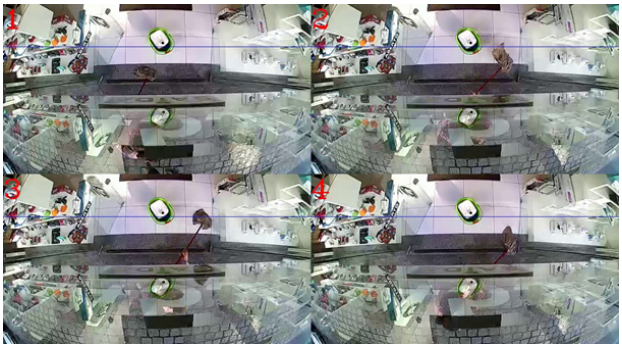

Fig. 2. Example using squeegee in the cleaning process

Table 4. Counting benchmark's metrics (cognitive approach) in the dynamic light scenario

Direction	Up			Down		
Metrics	Precision	Recall	F1score	Precision	Recall	F1score
Video 1	0.6499	0.9999	0.7878	0.6190	0.9999	0.7647
Video 2	0.7499	0.9999	0.8571	0.6666	0.9999	0.7999
Video 3	0.9999	0.8999	0.9473	0.9999	0.7777	0.8749
Video 4	0.7499	0.8571	0.7999	0.7999	0.9999	0.8888
Comb.	0.7560	0.9393	0.8378	0.7317	0.9374	0.8219

Running the same benchmark but with the cognitive approach the F1score, Table 4, improved more than 10%, improvement more visible in the first video where the cleaning squeegee was ignored by YOLO.

6.2 Men and women counting performance

A camera was installed in front of the door to capture a sequence of frames (20 frames were saved for each entrance confirmation), Fig. 3.

Fig. 3. Sequence of frames example of a person walking in the perfect scenario

The trigger to save the frames is the YOLO-based counter and the Table 5 contain the metrics extracted. It was observed that most the times the CNN committed a mistake the cause was occlusions and other people in the door space causing detections and tracking problems.

Table 5. Gender classification benchmark's metrics in the perfect scenario

Direction	Up			Down		
Metrics	Precision	Recall	F1score	Precision	Recall	F1score
Video 1	0.6249	0.7692	0.6896	0.9999	0.9999	0.9999
Video 2	0.3846	0.7142	0.4999	0	0	0
Video 3	0.9999	0.7499	0.8571	0.7999	0.7999	0.7999
Video 4	0.9999	0.7499	0.8571	0.9999	0.9999	0.9999
Comb.	0.5999	0.7499	0.6666	0.7777	0.5384	0.6363

The benchmark done in a real store, metrics in Table 6, achieved a very low performance. As the camera was supposed to be used for other reason, it was installed far away from the door, Fig. 4, making it difficult for the CNN to detect and classify. Although YOLO detected the person as the features were poor it could not resolve to a good classification exposing a tendency to classify as a male.

Table 6. Gender classification benchmark's metrics in the real scenario

Direction	Up			Down		
Metrics	Precision	Recall	F1score	Precision	Recall	F1score
Video 1	0.2499	0.2499	0.2499	0	0	0
Video 2	0	0	0	0	0	0
Video 3	0.7499	0.3749	0.4999	0	0	0
Video 4	0.5714	0.7999	0.6666	0	0	0
Comb.	0.4499	0.4090	0.4285	0	0	0

Fig. 4. Sequence of frames example of a person walking in the real scenario

7 Discussion

As observed in the benchmarks people walking in were counted more accurately, therefore that should be the information shown to the store's owner. Even though it demonstrated potential, the traditional-based counter achieved a F1score result of roughly 0.7 not enough for commercial use. On the other hand, YOLO-based counter achieved a F1score better than 0.8, making its information much closer to the real world.

In good and perfect conditions of light and distance to the door the gender classification module achieved a F1score of 0.66 (male) and 0.63 (female) also demonstrating potential, but not being good enough for commercial use. It is important to point that the benchmark tested the whole system and some errors were counting people module's fault.

The videos in the university laboratory were record with everybody consensus and the videos in the real store had the owner's approval, the images were used for running benchmarks only.

8 Conclusion

The disadvantages physical stores face when compared with virtual stores to obtain client's information was approached in this paper. The fact that virtual stores have better understanding of its clients, strategies will be easier and more efficiently executed on the sales process.

Even though the computational tool could not achieve the expected performance for commercial use, it demonstrated that it has potential and there is space for optimization and refinement. The counting people module using background subtraction approach needs to avoid the false positives and a possible path is to classify the detections as a person or not.

To improve the gender classification module, it could be done by improving the neural network, if it achieves a very high F1score the classification on a sequence of frames shouldn't be necessary, but only the one on the instant the person walks in. Also, as discussed on section 4 the dataset used to train the current system suffers from some

missing annotations and by correcting the dataset and retraining the network the performance should improve.

9 References

1. Hwangbo, H., Kim, Y. S., & Cha, K. J. (2017). Use of the smart store for persuasive marketing and immersive customer experiences: A case study of Korean apparel enterprise. Mobile Information Systems, 2017.
2. Quintana, M., Menéndez, J. M., Alvarez, F., & Lopez, J. P. (2016). Improving retail efficiency through sensing technologies: A survey. Pattern Recognition Letters, 81, 3-10.
3. Maria Alice V. Rocha, Lynne Hammond, David Hawkins, (2005) "Age, gender and national factors in fashion consumption", Journal of Fashion Marketing and Management: An International Journal, Vol. 9 Issue: 4, pp.380-390.
4. Jayaram, Athul. (2017). Smart Retail 4.0 IoT Consumer Retailer Model for Retail Intelligence and Strategic Marketing of In-store Products.
5. Karim, N. T., Jain, S., Moonrinta, J., Dailey, M. N., & Ekpanyapong, M. (2018, January). Customer and target individual face analysis for retail analytics. In 2018 International Workshop on Advanced Image Technology (IWAIT) (pp. 1-4). IEEE.
6. S. Jeeva, M. Sivabalakrishnan, Survey on Background Modeling and Foreground Detection for Real Time Video Surveillance, Procedia Computer Science, Volume 50, 2015, Pages 566-571.
7. Desai, H. M., & V. G. (2014). A Survey: Background Subtraction Techniques. International Journal of Scientific & Engineering Research, 5(12), 1365-1367.
8. K.T, T.T., & T. Z. (2011). Making Background Subtraction Robust to Various Illumination Changes. International Journal of Computer Science and Network Security, 11 (3), 241-248.
9. Zeiler M.D., Fergus R. (2014) Visualizing and Understanding Convolutional Networks. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham
10. S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.
11. Redmon, Joseph & Divvala, Santosh & Girshick, Ross & Farhadi, Ali. (2016). You Only Look Once: Unified, Real-Time Object Detection. 779-788. 10.1109/CVPR.2016.91.
12. J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6517-6525.
13. Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Springer, Berlin, Heidelberg.