Image search engine for furniture recommendations

Mariana ȘERBAN-PREDA
The University of Economic Studies, Romania
serbanpredamariana18@stud.ase.ro

Artificial intelligence (AI) has made enormous strides in recent years, transitioning from science fiction to a technology that is revolutionizing every sector of the global economy. Thanks to advances in machine learning, natural language processing and computer vision, artificial intelligence is no longer a futuristic dream, but a reality today. Once optimised and integrated into everyday life, artificial intelligence will substantially enhance human capabilities and contribute to the betterment of society. This paper will embrace the opportunities offered by computer vision, as part of artificial intelligence, by showcasing the FurnishMe application, an image search engine for furniture recommendations. Buying furniture online, as well as offline, is an overwhelming process given the quantity and diversity of furniture products available. The FurnishMe software solution allows users to easily explore the furniture of the three largest furniture retailers in Romania: Ikea, Jysk, and Dedeman. The system analyses user-uploaded interior design images in order to identify furniture items and provide aesthetically similar products from the three big retailers mentioned above. Both consumers and traders benefit from this solution. Clients benefit from a quick and easy way to choose the product they desire which integrates various features such as design, texture and colour. Moreover, businesses gain from greater sales by luring clients and saving time on in-person consultations.

***Keywords**: Deep Learning, Convolutional Neural Networks, Object Detection, Furniture recommendations, Web Scraping, Computer Vision, Visual Search Engine*

1 Introduction

In recent years, the technology innovation has helped propel e-commerce forward, making online purchases of furniture increasingly commonplace. As civilization has evolved and living standards have increased, people have inclined to establish visual comfort in their houses in addition to safety and physical comfort. As part of an increasingly digital environment, it's simply natural for technology developers to search for ways of providing an easy and personalized shopping experience when purchasing furniture online. According to Ozturkan [1], 8 out of 10 consumers who have been impressed by giants like IKEA or Amazon have bought furniture online thanks to hyper-personalised marketing, augmented reality and the ability to contextualise potential purchases in their own home.

E-commerce and the furniture industry have experienced impressive growth over recent years. However, searching for and choosing furniture online can be a challenging and time-consuming activity which decreases user satisfaction through poor choices or limited options available.

Furniture shopping can be an overwhelming task for the majority due to the variety of furniture products on the market. Moreover, searching through specialized website catalogues is a time-consuming activity since there are no compatibility checks between textures, colours or dimensions of the products. Consumers are frequently inspired by interior design images on social media, yet they may find it challenging to discover specific or similar furniture products. This step requires browsing through multiple online stores for products with matching features.

The aim of this study is to create a computer vision technology-based information system that simplifies the home furnishing process. Based on the recognition of furniture objects, the system will generate relevant

and personalised product recommendations from partner furniture stores. As a result, users will benefit from an improved experience in finding the right furniture for their needs and preferences.

In the next sections of this paper, we will present methods, technologies and results obtained during this research project in order to demonstrate its efficiency and usefulness for furniture recognition and recommendation.

2. Literature Review

The information technology system to be implemented is an image-based search engine that stands out from other similar systems by integrating a furniture object detection module.

Today, computer vision-based information systems provide deep and truthful analysis of visual reality thanks to deep learning techniques and convolutional neural networks (CNNs). These techniques have been applied in various fields such as fashion, medicine, transport or furniture. In the furniture industry, image-based recommendation engines can identify similar products based on design, style and various characteristics such as colour or texture.

It is essential to understand the accuracy and applicability of these techniques and to highlight new research directions.

In the paper "*DeepStyle: Multimodal Search Engine for Fashion and Interior Design"*, Trzcinski and his collaborators developed a multimodal search engine that combines visual and textual cues to retrieve elements aesthetically similar to the query from a multimedia database [2].

They aim to enable intuitive retrieval of products for fashion and interior design domains, as well as to address the inadequacies of currently existing search engines, by using a neural network architecture to model the joint multimodal space of database objects.

The search engine accepts two types of query information: an image that contains object(s) and a textual query that specifies the search criteria. YOLO detection model was used in order to detect objects of interest. Visual search identifies the image query's nearest neighbours in the space of extracted features. Textual queries limit the set of results to those that are relevant to the query.

The article discusses several neural network architectures used for multimodal search, including convolutional neural networks (CNNs), Siamese Networks, LSTM networks, and structure-content neural language models (SC-NLMs). The authors propose their own architecture called *DeepStyle* which uses a combination of these elements [2].

Another relevant study to this research presents a new interior style detection approach that uses multi-scale features and boosting to increase accuracy over traditional and residual network methods. With this approach, interior style detection involves non-hierarchical clustering and multi-scale feature fusion by using spatial pyramid matching (SPM), colour information and object detection into bag of visual words (BoVW) [3]. Features extracted from images may include colour histogram, vector colour analysis and local feature histograms. The algorithm also involves predetermining cluster number in advance, training with rule-based boosting and using LightGBM to estimate room style. The proposed method outperforms conventional BoVW methods and residual network (ResNet) in terms of accuracy.

The literature has highlighted various approaches to the development of image-based search engines. Considering the detailed notions previously discussed, the primary objective of this paper is to integrate advanced image processing algorithms into an efficient and relevant solution for the needs of the furniture industry and consumers.

p_c	b_x	b_y	b_w	b_h	c_1	c_2	...	...	c_n

3. Methodology

The architecture of the proposed solution is composed of several interconnected components that work together to provide functionalities like object detection and furniture recommendation. Figure 1 shows a diagram of the solution architecture:

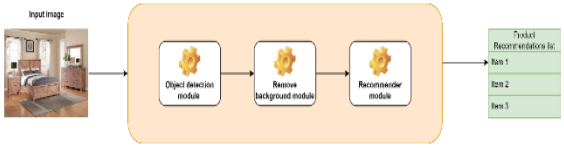

Figure 1. Solution arhitecture

A. Object Detection Module

YOLOv5 neural network was trained to detect pieces of furniture. You Only Look Once (YOLO) is a state-of-the-art algorithm for real-time object detection. Outstanding features for this algorithm are speed and accuracy in detection. YOLO uses a single neural network to perform classification as well as calculating the coordinates of the bounding boxes of objects in an image. It is currently an open-source algorithm that benefits from the support of a large community that constantly contributes to its improvement [4].

The first step in object detection using the YOLO algorithm starts by dividing the original image into NxN cells of equal shapes. Thus, each cell is responsible for localizing and classifying the object that it covers, along with a confidence score. This confidence score reflects the model's certainty that an object exists in that cell and that the delineation is accurate [4,5].

The next step is to determine the bounding boxes. YOLO determines the values for the bounding boxes' attributes using a regression algorithm where Y is the final vector representation for each bounding box and has the following structure [4]:

- p_c represents the probability score of the cell containing an object;
- (b_x, b_y) indicates the centre of the predicted bounding box;
- (b_w, b_h) represents the size of the bounding box;
- c_i corresponds to the class of objects predicted by the classification algorithm

Frequently, there may be multiple bounding box predictions for a single object. Using the Intersection over Union (IoU) and Non-Max Suppression metrics we manage to eliminate redundant detections. IoU measures the overlap between two bounding boxes which are usually the predicted box and the ground truth box. The metric has a range from 0 to 1 and is defined as the ratio of the area of intersection to the area of meeting of two rectangles as shown in Figure 2. The higher the value of IoU, the more the predicted box overlaps with the ground truth box. To eliminate redundant boxes, the NMS metric compares the previously calculated scores and eliminates those boxes with lower scores to the advantage of overlapping boxes with a higher score [6].

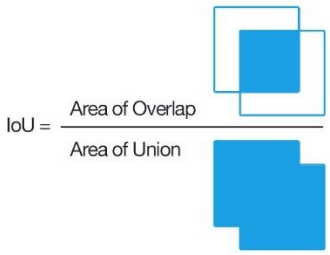

Figure 2. Intersection over union

Architecturally, the YOLO model consists of three key components: head, neck and backbone as shown in Figure 3. The backbone component is composed of convolutional layers. It is responsible for extracting features from an input image. Firstly, it is trained on a classification dataset, such as ImageNet. Also, the backbone component is usually trained with a lower resolution than the final detection model, since detection requires finer details than classification [5]. The neck component uses the features of convolution layers in the fully connected layered backbone to make predictions about the probabilities and bounding box coordinates. The head component is the final output layer of the network that can be interleaved with other

layers with the same input shape for transfer learning. These three portions of the model work together for feature extraction, object classification and localization [5,7].

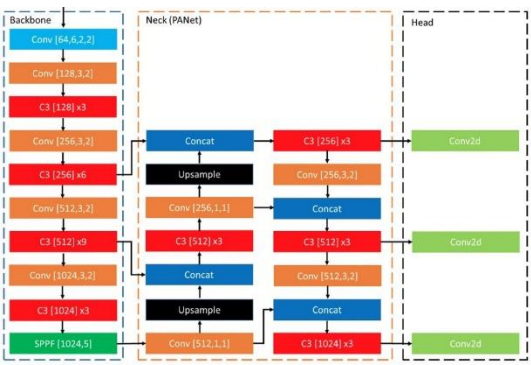

Figure 3. YOLO architecture [7]

B. Background removal

Background removal is a key technique which improves the recommendations of furniture products. The only relevant features for finding similar objects are the features of the detected object and not the environment. Consequently, we used Mask R-CNN model to segment the detected object and to translate it into an image with a white background. Thus, by reducing the background noise we manage to improve the accuracy of the recommendation algorithm. Due to the lack of a dataset including the object masks of the FurnishMe application's eight classes, we used the pre-trained model, provided by the open-source platform Detectron2, whose modules could be used by simply installing them in the Anaconda virtual environment.

C. Integrating with Ikea, Jysk, Dedeman store websites

The web scraping process allows the application to be integrated with the most popular e-commerce sites in the furniture industry. The Selenium library was used to browse the websites of IKEA, Jysk, and Dedeman stores, while Beautiful Soup, a library for parsing HTML and XML documents, was used to extract data. Images of products linked to the eight categories were downloaded and saved to disk, and details such as name, price, and external link were saved in the database.

D. Reverse Image Search Engine

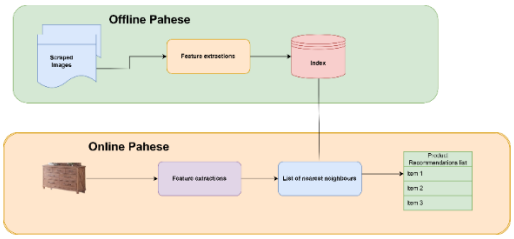

Figure 4. Image Search Engine architecture

In terms of recommending furniture products from the websites of Romania's most popular retailers, our primary goal was to develop a module that relies solely on visual similarity. It was identified the need to transform the information contained in millions of pixels of an image into a small representation that can be compared with other representations. A convolutional neural network takes an image as input and converts it into a feature vector. The latter serves as input to a classification layer which identifies the class label [2]. An ideal way to find similar images would be to pass the images through a high-performance convolutional neural network, extract the features and use a metric such as Euclidean distance or angular cosine to calculate the closest feature vectors.

Resnet neural network was used to understand the distinctive features of the images and represent them as feature vectors. Similar images are retrieved based on metrics such as Euclidean distance or angular cosine. A major advantage is that Resnet18 is available as a pre-trained model so that it can be used for image recognition without being trained from scratch.

Architecturally, Resnet is composed of residual blocks that allow easy propagation of information through the network. By its structure the algorithm solved the problem of information loss in a multi-layer neural network [8]. This phenomenon is called the gradient vanishing problem. In the process of back propagation, the gradient (a measure of the sensitivity of the neural network to changes in its weights) shrinks, causing the initial layers not to train properly which would lead to poor network performance [9].

As a result, the residual blocks contain skip connections that allow information to reach the previous layers.
Images downloaded from external sites were passed through the ResNet-18 convolutional neural network to extract features. Feature vectors were indexed thanks to the Spotify Annoy library which allows approximate nearest neighbour search for points in n-dimensional space.
In order to extract feature vectors, the model was modified by removing the last layer which was responsible for classification. It was provided with 512 features, identified as inputs by the last convolutional layer, and it used them the to identify one of the 1000 possible classes. Thus, the 512 features will be provided to the index which based on the angular cosine metric will be able to identify the N nearest neighbours
Index creation is an independent process that can be done only once and shared between different resources. Any image uploaded within the FurnishMe software solution is transformed into a feature vector using the ResNet-18 neural network. The loaded index provides a list of nearest neighbours using the angular cosine method.

4. Experiment & Results

For the development of the furniture object detection model, a series of steps were taken, such as data collection, data processing, training object detection model and prediction. The original YOLOv5 deep learning model can be used free of charge in a pre-trained version on the MS COCO dataset. The MS COCO (Microsoft Common Objects in Context) dataset is widely used for detecting objects belonging to 80 different classes. The FurnishMe application aims to detect furniture objects belonging to 8 classes: bed, sofa, chair, table, curtain, mirror, lamp, cabinetry. The dataset mentioned above includes only half of the classes needed by the computer system (bed, sofa, table, chair). Therefore, the YOLO neural network requires an explicit training process on a custom dataset.

Training a model from scratch is a difficult and time-consuming task. Therefore, I chose transfer learning method which reuses a previously learned model for solving a similar new task. In transfer learning, a machine exploits knowledge acquired from a previous task to improve the generalisation of a new one. In computer vision, neural networks are built in layers that progressively identify features such as lines, colours, textures or shapes, finally reaching the identification of specific features for classification. In the transfer learning technique, layers that identify common features are reused and those used for classification are retrained [10].
Data preparation is another complex and time-consuming process, because YOLO neural network requires the training dataset to be defined in a specific format. For each image in the dataset, a text file with the same name is created. Each text file contains the annotations on the image's item bounding boxes. The structure of a line in the text file consists of:

- object class - an integer representing the object class code. The class index must start at 0 and increase by 1 for each unique class in the dataset;
- X_{centre}, Y_{centre} the coordinates of the centre of the bounding box normalised to the width and height of the image;
- W, H - length and height of the bounding box normalised to image length and height

The image annotation process can be done manually using specialised software such as LabelImg. In this paper, images of interest were taken from a public source of labelled data, Open Images.
A training dataset (80% of the obtained images) and a test dataset (20% of the retrieved images) were created. The training set is used to teach the model how to detect furniture objects. The test dataset is used to assess the model's ability to generalise and predict on unknown data that was not part of the model training process.
The Google Collaboratory service was chosen to train the model thanks to its GPU

which speeds up the training time of a deep learning model. The model was trained for 100 epochs with the following results:

```
Validating runs/train/Model/weights/best.pt...
Fusing layers...
Model summary: 157 layers, 7031701 parameters, 0 gradients
                 Class     Images  Instances          P          R      mAP50   mAP50-95: 100% 42/42 [00:21<00:00,  1.99it/s]
                   all       2659       8468      0.573      0.545      0.514      0.312
                   Bed       2659        488      0.752      0.709      0.738      0.471
                 Couch       2659        491      0.586      0.615      0.589      0.353
                 Chair       2659       2930      0.495      0.505      0.458      0.244
                 Table       2659       1512      0.477      0.474      0.484      0.214
            Cabinetry       2659       1181      0.383      0.355      0.263      0.137
                  Lamp       2659        789      0.582      0.449      0.439      0.232
               Curtain       2659        753      0.534      0.565        0.5       0.29
                Mirror       2659        324       0.77      0.688      0.718      0.554
Results saved to runs/train/Model
```

Figure 5. YOLO training results

Figure 5 shows the performance metrics. The Precision (P) is represented by the proportion of correct positive predictions relative to all positive predictions (correct or incorrect) made by the model [11]. This metric is useful if we want to reduce the number of false positives, i.e. cases where the model predicts that an object is in the image but in fact it does not exist.

The Recall (R) metric is the ratio of the number of correct positive classifications to the sum of the number of correct positive predictions and the number of incorrect negative predictions [11]. This metric is useful when we want to minimise the number of false negatives, i.e. cases where the model does not identify an object in the image even though it exists [12]. Thus, the P metric shows the accuracy in classifying objects as positive, and the R metric measures the ability of the model to detect positive objects.

According to the Figure 5, the model recognizes an object properly in 58% of cases and it identifies correctly the label of all furniture objects in 55% of cases. For the IoU (Intersection over Union) metric, it is necessary to choose the acceptable threshold for which the model prediction is positive and correct. Usually, the accepted threshold is 50%. By progressively calculating the Precision and Recall metrics for different thresholds, we obtain the Precision-Recall curve visible in Figure 6. We thus observe that as the IoU threshold is lower, the Recall metric which show the ability to detect positive objects increases, but the accuracy decreases.

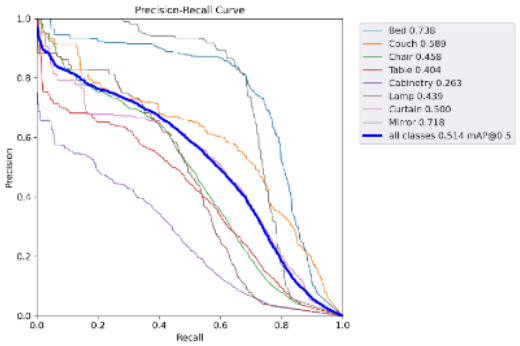

Figure 6. Precision-Recall Curve

Mean Average Precision (mAP) is another important metric in evaluating an object detection model. This metric is a summary of the Precision-Recall curve into a single value that represents the average of all accuracies across classes [11].

Overall, the model exhibits average performance, with moderate metric values for most classes. We note the increased accuracy for labels “bed” and “mirror”. The mAP50 score indicates the model's ability to detect objects consistently with an overlap of more than 0.5 and with an accuracy of about 52%. However, these scores could be improved by optimising the model or increasing the training dataset.

To identify possible errors in the classification process, we analyse the confusion matrix shown in Figure 7.

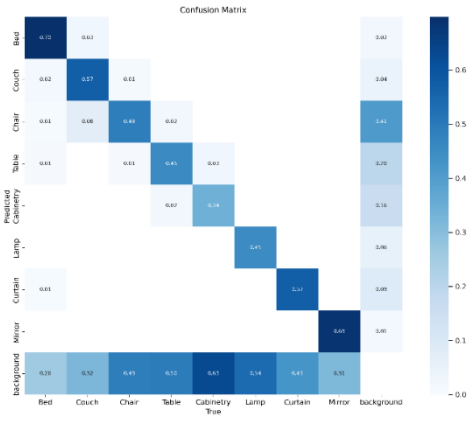

Figure 7. Confusion matrix

The classification accuracy for each class is shown on the diagonal of the confusion matrix. The values range from the minimum accuracy value, recorded for the cabinetry class (0.34), to the maximum accuracy

value, recorded for the bed class (0.7). The confusion matrix reveals a low level of misclassification between furniture items. For example, in some cases the sofa is classified as a chair. The tendency to categorise objects from the classes of interest to the background class has been highlighted as a drawback in the model. This may signal a lack of balanced training data or inadequate architecture.

The purpose of the FunirshMe software solution is to demonstrate the feasibility of a concept by developing a prototype software solution. Due to the limited open-source resources and the model's average performance with moderate metric values, we chose to use it

Since the ResNet-18 neural network was only used for feature vector extraction (and not for classification), the pre-trained version was chosen. In this way, we obtained meaningful image representations without engaging in extensive training. Such a neural network stands out for its adaptability thanks to training on a diverse dataset.

The Mask R-CNN model was used for instance segmentation process. Due to the lack of a dataset including the object masks of the FurnishMe application's eight classes, we used the pre-trained model.

Figure 8 shows the integration of the modules described above in a Flask web application.

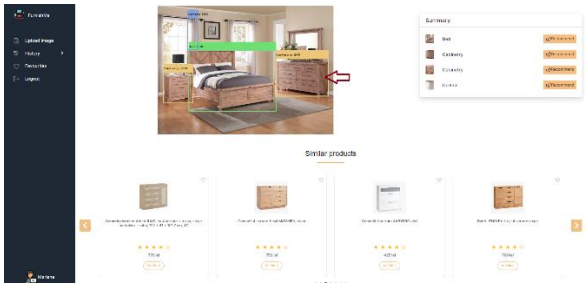

Figure 8. FurnishMe application

5. Conclusions

The main aim of this paper was to develop an innovative information technology system capable of providing users a personalised and easy home design experience. The solution is based on computer vision technology and it allows the users to explore the furniture of some of the most popular retailers in Romania: Ikea, Jysk and Dedeman. FurnishMe detects furniture items in user-uploaded pictures and recommends visually similar products from the three big shops mentioned above. The application adds value to the field under review by providing an efficient and convenient way for consumers to choose the right furniture. The software solution does not restrict itself to the products of a single retailer, but rather presents the products of Romania's most popular and accessible retailers. This aspect allows the users to analyse and evaluate products based on their specific needs, such as pricing, personal preferences, product availability in a particular store location and so on.

At the moment, the solution is only a prototype that can be extended and improved through additional research. The primary goal of this study was to demonstrate the feasibility of an idea. One aspect that would add value to the software solution is optimising the detection and recommendation algorithm's performance using various model regularisation techniques. Furthermore, user loyalty for the FurnishMe software solution could be strengthened by collecting user preferences and generating personalised recommendations.

Appendix

- CNN – Convolutional Neural Network
- YOLO – You Only Look Once (Object Detection Model)
- NMS – Non-Maximum Suppression
- IoU – Intersection Over Union
- ResNet – Residual Neural Network
- SC-NLMs - Structure-Content Neural Language Models
- LSTM – Long short-term memory
- SPM – Spatial Pyramid Matching
- BoVW – Bag of Visual Words
- LightGBM – Light Gradient Boosting Machine

References

[1] S. Ozturkcan, "Service innovation: Using augmented reality in the IKEA Place app," Journal of Information Technology Teaching Cases, vol. 11, no. 1, pp. 8-13, 2021.
I. Tautkute, T. Trzciński, A. P. Skorupa, Ł. Brocki, and K. Marasek, "Deepstyle: Multimodal search engine for fashion and interior design," IEEE Access, vol. 7, pp. 84613-84628, 2019.

[2] A. Yaguchi, K. Ono, E. Makihara, N. Ikushima, and T. Nakayama, "Multi-Scale Feature Fusion for Interior Style Detection," Applied Sciences, vol. 12, no. 19, pp. 9761, 2022.

[3] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: Challenges, architectural successors, datasets and applications," Multimedia Tools and Applications, vol. 82, no. 6, pp. 9243-9275, 2023.

[4] H. Liu, F. Sun, J. Gu, and L. Deng, "Sf-yolov5: A lightweight small object detection algorithm based on improved feature fusion mode," Sensors, vol. 22, no. 15, pp. 5817, 2022.

[5] X. Wang and J. Song, "ICIoU: Improved loss based on complete intersection over union for bounding box regression," IEEE Access, vol. 9, pp. 105686-105695, 2021. [7] T. N. Pham, V. H. Nguyen, and J. H. Huh, "Integration of improved YOLOv5 for face mask detector and auto-labeling to generate dataset for fighting against COVID-19," The Journal of Supercomputing, pp. 1-27, 2023.

[6] E. Limonova, D. Alfonso, D. Nikolaev, and V. V. Arlazarov, "ResNet-like architecture with low hardware requirements," in 2020 25th International Conference on Pattern Recognition (ICPR), pp. 6204-6211, Jan. 2021.

[7] D. Zou, Y. Cao, D. Zhou, and Q. Gu, "Gradient descent optimizes over-parameterized deep ReLU networks," Machine Learning, vol. 109, pp. 467-492, 2020.

[8] G. Pinto, Z. Wang, A. Roy, T. Hong, and A. Capozzoli, "Transfer learning for smart buildings: A critical review of algorithms, applications, and future perspectives," Advances in Applied Energy, pp. 100084, 2022.

[9] R. Padilla, W. L. Passos, T. L. Dias, S. L. Netto, and E. A. Da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," Electronics, vol. 10, no. 3, pp. 279, 2021

[10] H. Zhu, H. Wei, B. Li, X. Yuan, and N. Kehtarnavaz, "A review of video object detection: Datasets, metrics and methods," Applied Sciences, vol. 10, no. 21, pp. 7834, 2020.

Mariana SERBAN-PREDA – student at Bucharest University of Economic Studies, attending E-Business Master, Bucharest, Romania; obtained a Bachelor's Degree in Economic Informatics in 2021; passionate about Artificial Intelligence, Big Data and Machine Learning;