Virtual Reality manuscript No.
(will be inserted by the editor)

An augmented reality application for improving shopping experience in large retail stores

Edmanuel Cruz · Sergio Orts-Escolano · Francisco Gomez-Donoso · Carlos Rizo · Jose Carlos Rangel · Higinio Mora · Miguel Cazorla

Received: date / Accepted: date

Abstract In several large retail stores, such as malls, sport or food stores, the customer often feels lost due to the difficulty in finding a product. Although these large stores usually have visual signs to guide customers towards specific products, sometimes these signs are also hard to find and are not updated. In this paper, we propose a system that jointly combines deep learning and augmented reality techniques to provide the customer with useful information. First, the proposed system learns the visual appearance of different areas in the store using a deep learning architecture. Then, customers can use their mobile devices to take a picture of the area where they are located within the store. Uploading this image to the system trained for image classification, we are able to identify the area where the customer is located. Then, using this information and novel augmented reality techniques, we provide information about the area where the customer is located: route to another area where a product is available, 3D product visualization, user location, analytics, etc. The system developed is able to successfully locate a user in an example store with 98% accuracy. The combination of deep learning systems together with augmented reality techniques shows promising results towards improving user experience in retail/commerce applications: branding, advance visualization, personalization, enhanced customer experience, etc.

Keywords Smart Shopping · Deep learning · Augmented Reality · Retail Stores · User Experience · Human-Computer Interaction · 3D visualization

All the authors are with
Instituto Universitario de Investigación Informática
Universidad de Alicante
Tel.: +34-965903400
Fax: +34-965903902
E-mail: miguel.cazorla@ua.es

1 Introduction

Key business leaders in the technology sector are currently committing to the trend of augmented reality, as is the case of the technology giants Apple, Google and Microsoft.

The current level of technological development is the result of several key factors such as hardware improvement and the increase in mobile devices with high computing capabilities. Moreover, recent trends such as Internet of the Things (IoT), Deep Learning (DL) and Augmented reality (AR) approaches have also contributed to the way technological solutions are used and developed. AR consists of adding virtual information to the physical world, allowing the user to enrich their environment perception. The basic goal of an AR system is to enhance the user's perception of an interaction with the real world through supplementing the real world with 3D virtual objects that appear to coexist in the same space as the real world Azuma et al.. Besides, industrial IoT and Machine Learning (ML) are two of the areas where the potential of AR can be visualized in the near future Pauly et al. [2015], Ahn et al. [2015], Ortiz-Catalan et al. [2016].

Todays global economy is driven by daily trading operations made by customers at stores Goldman [2001]. Customers are permanently seeking to make the most of their incomes and so compare not only product quality but also the complete purchase experience Lemon and Verhoef [2016], Blazquez [2014]. Thus, stores must be adaptive to this type of consumer. Knowledge about products is a crucial aspect that increasingly interests consumers. This knowledge may be accessed by digital means Willems et al. [2017]. Nevertheless, the digital experience has not yet been carefully considered by commercial stores. Therefore, we aim to apply AR techniques to enrich user experience when shopping for a product in a physical store.

Commercial stores are one of the sectors with the greatest possibilities to implement AR in the short-term. Therefore, this work proposes the development of an AR mobile application that initially allows to determine the location inside a commercial store based on visual information coming from a monocular camera (color image). These images are captured using the mobile phone camera. Next, the mobile application guides the user to the location of a desired product as well as show a virtual version of the product and related information.

The main intent of this research work is to improve shopping experience using current ML and AR techniques. In this direction, we proposed a system that enables shopping experience by showing AR models to the users directly at the store, so products that are boxed or not available can be seen as if the were physically there. Besides, the system provides user navigation, improving the localization of certain products at the store, this feature can be extremely convenient at large stores/warehouses.

The rest of the paper is organized as follows: First, in Section 2 the state-of the art in the field is presented. Next, Section 3 details a description of the proposal. This is followed by the Section 4, where the procedures for testing

the proposed approach are described and the results of the experiments are presented. Finally, Section 6 includes the discussion and conclusions of the work.

2 State of the art

Grewal et al. [2017] predicts that AR is one of the emerging applications that will define the future of retailing. Thanks to the use of this technology, the customer decision-making process will be improved, including interactive visualization of products and advertisement.

The use of AR inside Large Commercial Areas (LCA) has not yet been exploited. This kind of areas needs a special attention for helping customers Likavec et al. [2008]. There exist different AR systems able to apply this technology to different scenarios Carmigniani et al. [2011]. Although there are several LCAs that have stands where a customer can find products, the use of a mobile device could improve the customer experience Veijalainen [2008]. A comprehensive review of mobile AR is found in Chatzopoulos et al. [2017]. A large gap in this area is the use of AR to guide the customer inside a LCA, although this has been used for navigation in outdoor applications Narzt et al. [2006].

The proposed system uses a scene classification method for user location. The scene classification or indoor place categorization problem may be defined as the problem of classifying an image as belonging to a scene category from a set of predefined labels Maron and Ratan [1998]. Scene classifiers are also helpful for specific tasks Martínez-Gómez et al. [2014] such as autonomous navigation, high-level planning, simultaneous location and mapping (SLAM), or human-robot interaction (HRI). image Scene classification is commonly addressed as a supervised classification process Wu et al. [2009], where input data correspond to perceptions, and classes to semantic scene categories. Current approaches are based on a two-stage building process: a) selecting the appropriate descriptors to be extracted from perceptions, and b) choosing a classification model to be able to deal with the extracted descriptors.

Relying on the use of images as the main perception mechanism, the descriptor generation problem is addressed by using computer vision techniques. In this process, the organization of the data extracted from the images plays an important role. This is presented clearly in two of the most widely used approaches: the Bag-of-Words (BoW) Csurka et al. [2004], Martinez-Gomez et al. [2015] and the spatial pyramid Lazebnik et al. [2006]. These two approaches allow the generation of fixed-dimensionality descriptors, required for most of the state-of-the-art classification models, built from any type of local features.

The use of DL techniques is considered a notable milestone in the research areas of computer vision and robotics LeCun et al. [2010]. DL provides classifiers capable not only of classifying data but also of automatically extracting

intermediate features. This technique has been applied to image tagging with surprising results.

For instance, since 2012 the winners of ImageNet competition Russakovsky et al. [2015] have use Convolutional Neural Networks(CNNs) Krizhevsky et al. [2012], for example AlexNet(2012) and Clarifai(2013). In addition to very large amounts of annotated data for training, DL requires high processing capabilities for classification. Recently in Rangel et al. [2016] it has been demonstrated that the use of DL techniques improves results in scene classification.

To the best of our knowledge, DL techniques have only been used for feature extraction and feature matching to improve current AR systems. In Akgul et al. [2016] a DL technique is used to improve the tracking of a given target for AR applications.

In this work, we propose a novel system that improves customer experience through the use of AR techniques. To do so, we utilize state-of-the-art artificial intelligence techniques that help create a smart experience through the use of an AR application. Thanks to the use of a DL-based classifier, we are able to accurately locate a customer within a large retail store. Based on this information, we can create a rich, customized experience that helps the customer through the shopping process.

3 System description

The system is composed of a mobile phone application and a web service that performs the classification task; it takes an image of an aisle and returns a label. Therefore, the mobile application will communicate with the server to obtain information related to where the user is positioned. The development of this system is divided in two phases, the first for collecting the data and training the classifier and the second for linking the trained classifier with a mobile device to test the proposed approach. Further details of these two main stages are provided below. Figure 1 shows an overall description of the proposed system.

3.1 User Localization using a DL-based classifier on a Remote Web Service

In order to suggest where to find a certain item, first we have to locate the user. To do so, the customer must use our mobile application to take a picture of the aisle where they are currently located and the application will detect their location to suggest the path to the item they are looking for. To locate a user from a picture of an aisle, we rely on CNNs. These kinds of networks are able to classify an input image based on their features and assign them a category.

Finally, the classification model is integrated in a web service in order to make it available to external clients. As making predictions with a CNN is computationally demanding, it is too costly to implement it on a mobile

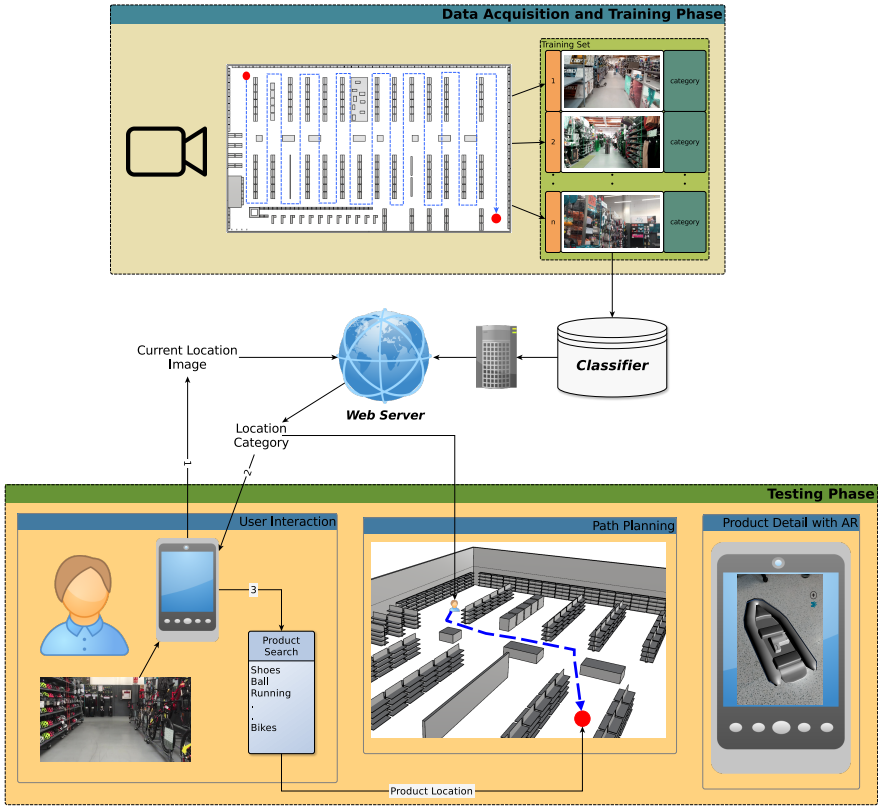

Fig. 1: Proposal flowchart. The user is located in (Bikes) and the item they are looking for is in (Golf). The path and location are computed by our mobile application.

device. Thus, when a user takes the image for location, the image is resized and sent to the web service. It classifies the image and returns the aisle where the user is located.

Next, we present the details of the user location subsystem.

3.1.1 Architecture

To locate a person in the store given a picture, we rely on a ResNet50He et al. [2015] Deep Convolutional Neural Network.

The ResNet50 DL architecture is currently the state-of-the-art CNN in image recognition, achieving a top-1 error (22.85%) on the ImageNet validation split. The main feature of this architecture is the inclusion of the "residual" term. It consists of the aggregation of the input image to the output image of a convolution block. As a result, the output of a convolution block may be seen

as the input image where the features activated by the filters are highlighted. In contrast, the output of a convolution layer in a default CNN is only the result of the neurons activation. If a neuron is not triggered in a certain region of the input image, the output remains with lower or null activation values. When the network computes the weights update in the backpropagation stage, the values in non-activated regions lead to very low upgrades, eventually even resulting in not upgrade at all, which causes the learning to block. This issue is known as the vanishing gradient problem Bengio et al., Pascanu et al. [2012]. The inclusion of the "residual" term helps fighting the vanishing gradient problem and allows the design of even deeper architectures. Currently, the best performer on several tasks of the ImageNet challenge is based on the "residual" approach introduced by ResNet.

Our proposal makes use of a default ResNet50, with a minor modification: the number of neurons was modified to 20 in the final fully connected layer in order to fit our problem.

3.1.2 Dataset and Training

In order to train the CNN, it is necessary to have labeled sequences of images belonging to the different aisles in the store. Hence, we recorded a video while passing through every aisle in a large sports store. The store was open at the time the dataset was recorded, so besides of the items, there are other customers in the store. This is desired feature because will force the learned model to deal with partially occluded scenes as we can not expect a full view of an aisles or shelves with no other customers in it. Three different human-operated RGB cameras were used in this process. The videos were split into clips and hand-labeled according to the aisle in which a clip was recorded. Then, the frames of every clip were extracted and resized to 224×224 in order to feed the CNN. Blurry frames were manually selected and excluded from the dataset. Table 1 shows the description of each class and the number of samples. Figure 2 shows a view of the environment (large retail store) where the proposed system was tested.

The samples were grouped by label, shuffled and split into training and test subsets with a 20% test split ratio.

The time investment on the dataset creation is distributed as follows: first, the time committed to film the aisles depends on the size of the store. As no special equipment (besides the RGB camera) or special requirements are needed, the time invested in this step is the time you spend wandering all the aisles of the store once. Then, the labeling process that consists on extracting the frames of the videos and assigning them a class is also a low time consuming task. As extracting the frames from a video is a fast and automatic task, we find the time of this task negligible. Also, assigning a category for each frame is also quite straight-forward as there should be a video per aisle. By far, the most time consuming is the model training process. This could take from several hours to a few days, depending on the hardware and the data. Nonetheless, this

Aisle	Description	# of samples
Transition	Corridors and communication paths	7419
Cycling	Cycling clothes and shoes	2682
Bikes	Bikes	1638
Fitness	Fitness exercise equipment	1341
Cardio	Cardio exercise equipment	1817
Sports Bras	Sports bras and women underwear	240
Woman Clothes	Assorted woman clothes	481
Martial arts	Kimonos, boxing gloves and punching bags	2182
Trekking	Trekking boots and equipment	2035
Ping-Pong	Ping-Pong blades, balls and tables	926
Bow and Lawn bowling	Lawn bowling equipment. Bows, arrows and targets	795
Football Shoes	Football shoes	604
Fishing	Fishing rods and baits	308
Hunting	Hunting equipment, camouflage clothes and baits	848
Football Children	Football equipment, shoes and socks for children	781
Golf	Golf sticks, bags and balls	760
Horse Riding	Horse riding boots, saddles and other equipment	818
Walking	Walking clothes, shoes and sticks	1184
Socks and Sports Shoes	Assorted sports shoes and socks	886
Street Sports	Roller blades, skate boards and scooters	254
Water Sports	Bathing suits and swimming goggles	306

Table 1: Description of the aisles that our localization system can detect.

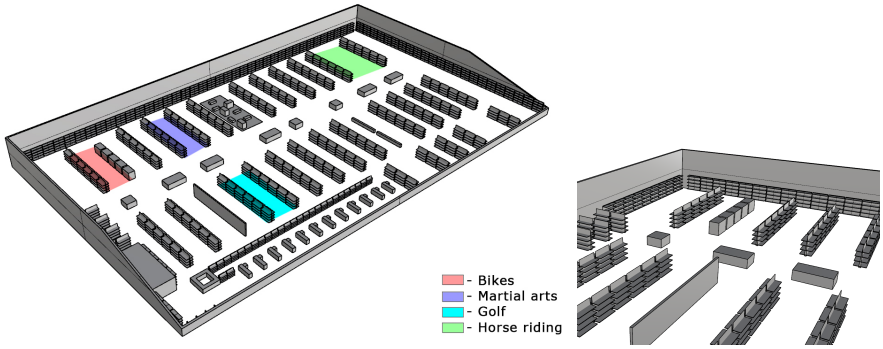

Fig. 2: Environment where the experiments were carried out. Only 4 of 20 categories are shown

is an automatic process and no personnel is involved but for the configuration of the parameters at the beginning of the training process.

Overall, we think that the investment of time is negligible given the potential value that the mobile application would provide.

3.1.3 Dealing with environment changes (visual appearance)

As a matter of fact, changes in the organization of the shelves, the modification of the packaging or products, or other may modify the visual appearance of the aisles. As the localization procedure relies on visual features, these events will make the system fail.

If there are slight changes in the environment (like changing a certain model of a shoe for another model), the deep neural network architecture is inherently able to deal with them. If there are major changes (like complete aisle reorganization or the addition of a new category of items), it is required to retrain the network. In this case, two methodologies for new data acquisition can be followed: if the major changes are gradually introduced, the users would provide enough training data when the localization fails and they manually choose the location, as explained in Section 3.3. If there are dramatic and sudden changes, a new data acquisition procedure is needed, so the affected aisled must be filmed. Either cases, the network must be retrained. Besides, we are considering implementing continuous learning techniques, where data acquired by the user is leveraged for incremental training (Learning without Forgetting Li and Hoiem [2016]).

3.2 Augmented Reality Mobile Application

In this work, we have developed an AR mobile application that utilizes DL techniques to provide a helpful experience when shopping in large stores.

Customers frequently feel lost when shopping in big stores. Thanks to the developed mobile application, users can easily locate themselves in the store, as explained in Section 3.1, and use mobile application guidance to find and retrieve information about the product they are looking for. Using AR techniques, we are also able to provide the user with a richer experience, being able to visualize a 3D model of different products available in the store. Thanks to current AR techniques, we are able to blend realistic off-line captured 3D models with the real world. This feature allows the user to know the availability of the different types or models of the products, e.g size and colors. It also helps companies reduce the amount of products customers return every year,as well avoids users having to remove the product from its packaging before buying it. According to the National Retail Federation (NRF) [1], Americans returned $260 billion in merchandise in 2015, a 66 percent increase from five years ago. The reason for these return is, in the majority of the cases, the lack of knowledge of the product by the final costumers. Therefore, using our mobile application, costumers have the option of knowing several features of the product before its purchase, avoiding the returning of the product to the store. Figure 3 shows a visual example of how AR is used in our application by displaying a 3D model of a product, and enabling the user to visualize

[1] https://nrf.com/sites/default/files/Images/Media Center/NRF Retail Return Fraud Final_0.pdf

the different customization choices of the item. Also, the user can rotate the model, perform zooming, change its features from the available pool or even check the in-store availability.

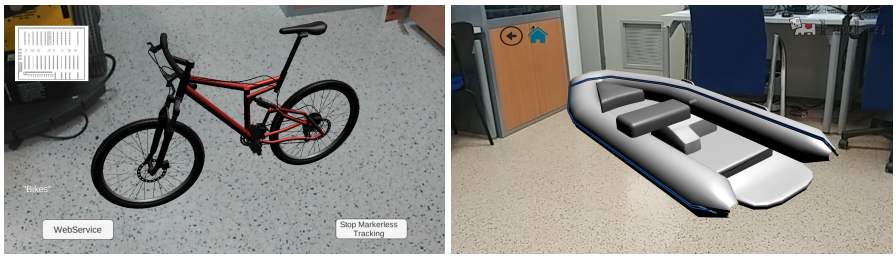

Fig. 3: Example of 3D models displayed in a store using AR techniques (Markerless tracking).

3.3 Mobile Application workflow

As stated before, our mobile application is able to automatically detect where the user is located and provide guidance to the item that the user is looking for. Furthermore, the user can project and manipulate virtual items in an AR fashion in order to check whether the item satisfies their personal preferences (such as size or color) or not.

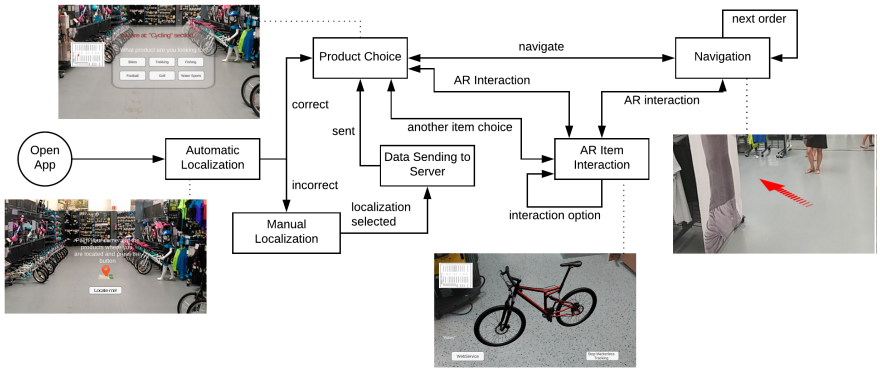

Fig. 4: Mobile Application workflow

The mobile application workflow, which is depicted on Figure 4, is as follows:

Once the users launch the mobile application, they must provide an image of the aisle where they are currently located using the mobile application. This image is sent to the web server and then the category for the image is predicted. The category is inferred from available aisles in the store taking advantage of DL methods, as explained in Section 3.1. The category assigned by the web server to the image determines the location of the user in the store.

The localization result is reflected on the mobile application UI and then, if it is correct, it asks the user what kind of product they are interested in. The mobile application features a complete searcher with filtering and different sorting option in order to enable a fast item searching process. Once an item is selected, the user can navigate to that object or project it in an AR fashion. The user can interact with the 3D model of the object displayed on the mobile device screen by rotating it, changing its color or singular features of each item available for purchasing. Then, the user can navigate to the item if the user chooses to. At all times, the user can query the in-store availability or customization options of the item.

The navigation screen projects an arrow in an AR manner, and behaves like a compass, pointing towards the next waypoint. Finally, if the users follow the directions, it will lead them to the location of the object they have chosen. In this point, the user could interact with a 3D model of the item, as explained before.

As stated earlier, the first step of the mobile application is the automatic localization of the user. In case that this process fails, as depicted in Figure 10, the user could manually localize itself in order to use the guidance capabilities of the mobile application. Since the automatic localization is performed from an image, if it is not accurate and the user must use the manual localization, both the image and the correct localization chosen by the user are sent to the server for further analysis.

As explained, our system mixes techniques from AR and DL to provide the user with a complete experience when visiting a store and purchasing a product.

4 Experiments

4.1 Experimental Scenario

The experiments in this work were carried out using the aforementioned dataset. Figure 5 shows representative images for 8 of the 20 available categories in the store and training dataset.

4.2 User Localization using Deep Learning

The architecture described in Section 3.1.1 was trained to take a picture of the users surroundings as an input and to infer which aisle appears in the

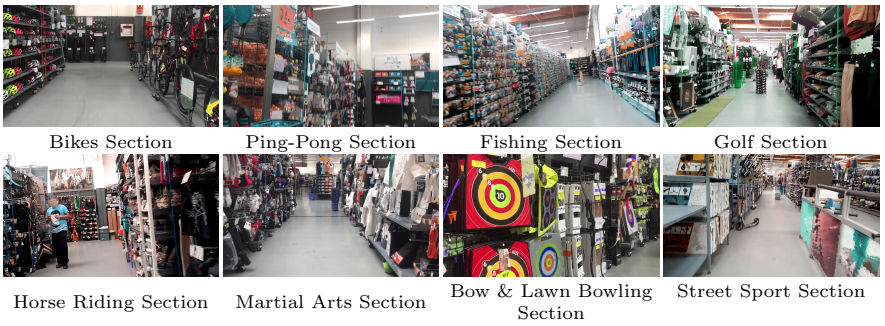

Fig. 5: Examples of 8 from the 20 different sections in the store.

input image. The details about the dataset we used for training and testing are shown in Section 3.1.2. As previously mentioned, the dataset was split in training and testing using an 80% and 20% ratio respectively. The optimizer of choice was Adam Kingma and Ba [2014] with a learning rate of 0.00001. Test loss and accuracy rates were used as an early stopping criteria. At the end of the training process, it achieved a 98.97% test accuracy. Figure 6 shows the confusion matrix related to this experiment.

The model generated is implemented as a web service to perform inference from remote devices. The system only takes 0.014 seconds at runtime, so it can be comfortably used in real-time applications. Consuming the web service from a mobile phone adds some latency. We measured this latency on different mobile phones and under different load conditions (WiFi, 3G/4G network) and, on average, it adds $\sim 1.5s$ of latency.

All timings, training procedures and results were obtained by conducting the experiments in the following test setup: Intel Core i5-3570 with 8 GB of Kingston HyperX 1333 MHz DDR3 RAM on an Asus P8H77-M PRO motherboard (Intel H77 chipset). Secondary storage was provided by a Seagate Desktop HDD 3 TB. Additionally, the system included a NVIDIA GTX1080Ti GPU used for training and inference.

The framework of choice was Keras 2.0.1 with Tensorflow 1.2.0 as the backbone. CUDA 8.0 and cuDNN were enabled.

The experiments were carried out using a real mobile phone device inside the store. Figure 7 (left) shows the mobile application interface for taking a picture of the current location in the store, asking the user to locate himself in the place.

Figure 7 (right) shows a product being displayed with AR for the virtual interaction with the customer inside the store.

The mobile application has been tested on several mobile devices with different hardware (graphics capabilities) specifications. The mobile application was developed using an AR library (Kudan[2]) which provides us with mark-

[2] https://www.kudan.eu/

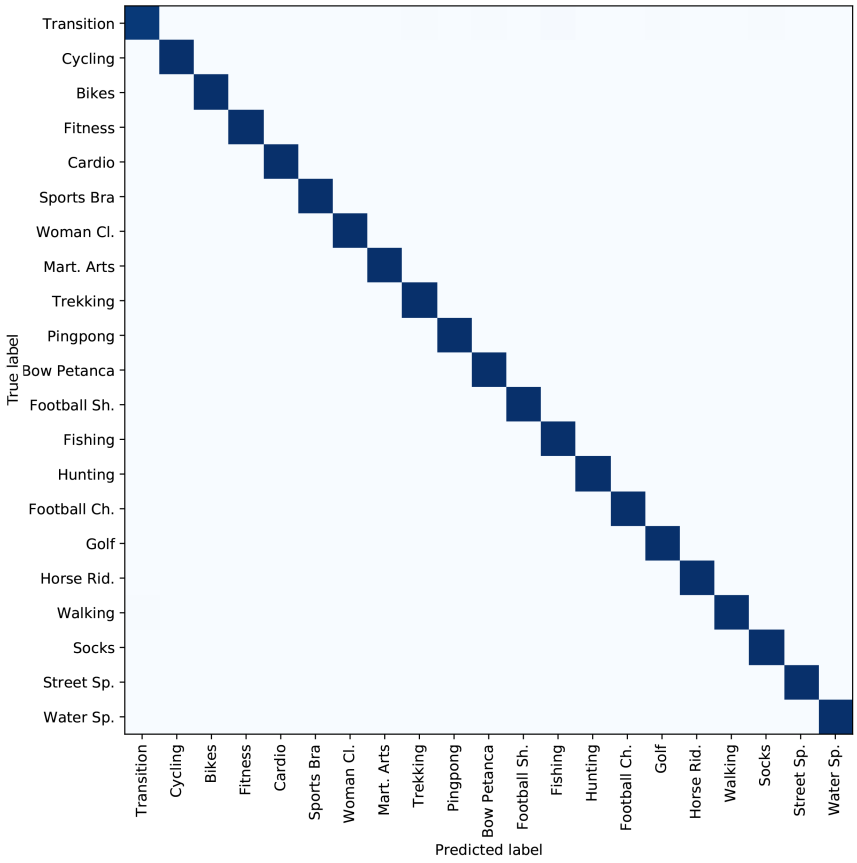

Fig. 6: Confusion matrix for the test split.

erless tracking of the scene Klein and Murray [2009], Newcombe et al. [2011]. The mobile application was developed using Unity3D and tested on mobile phones based on the Android and iOS operating systems.

4.3 Results

We tested the proposed mobile application in a real scenario. Figure 8 shows several examples in which the mobile application provides the category where the customer is located.

The mobile application developed also provides the user with guidance, being able to guide them to the aisle where the product they are looking for is located. Instructions are provided using AR techniques on the store floor. Figure 9 shows an example of the system developed for customer guidance. Guidance behavior is similar to the way a compass works, the arrow turns

Fig. 7: Mobile application screen captures.

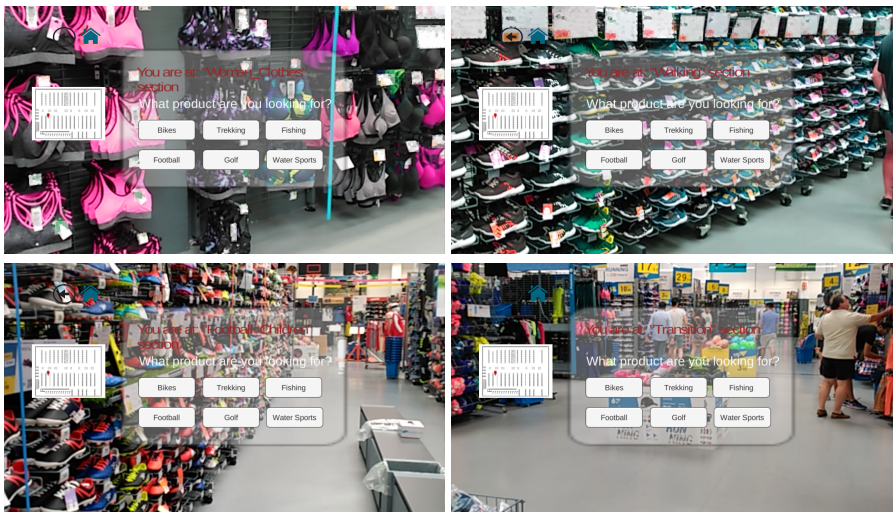

Fig. 8: Several results obtained by the proposed mobile application in a real scenario.

looking in the direction that the customer has to follow to find the right path as explained in Section 3.3.

5 Discussion

Having an Internet connection is one of the main limitations of the presented work. Since the use of DL techniques require high-performance computing to process input data in real-time, it is not possible to perform the image recognition process on the mobile phone. Besides, it would dramatically increase power consumption. The second limitation of the presented system is related to ML part of the presented system. Since the DL model was trained using product and store images acquired at an specific time, if the visual appearance of these products change, the system would need to be re-trained with

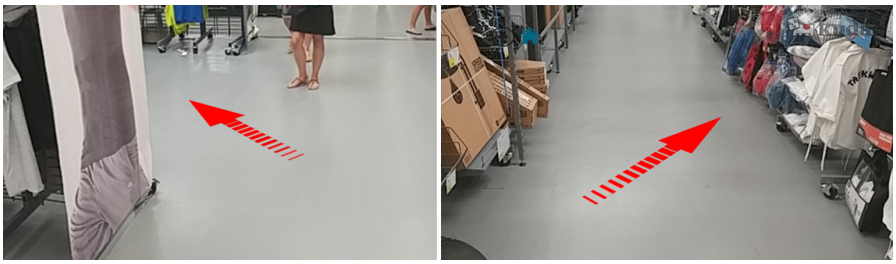

Fig. 9: Guidance example.

new acquired images. However, this is a problem that is being currently studied and there already exists multiple approaches to perform Learning without forgetting Li and Hoiem [2016]. In this way, the DL technique continuously retrained adding new data without forgetting previous knowledge.

We have detected that our system has some failure cases. Figure 10 shows an example of this. These cases occur when the user takes a picture and the system provides a different localization (corridor) from the real one. In that case, the system is unable to guide the customer correctly. We notice that this situation only happens 1 out of 20 predictions. We mitigate this error including the following: when the predicted localization is provided to the customer, he/she is asked if that is correct. If the localization is wrong and the customer confirm the localization, the system would guide to an incorrect destination. But, if the customer knows his/her correct localization and could rectify it, then the system uses that information to modify the learned model and improve the future predictions.

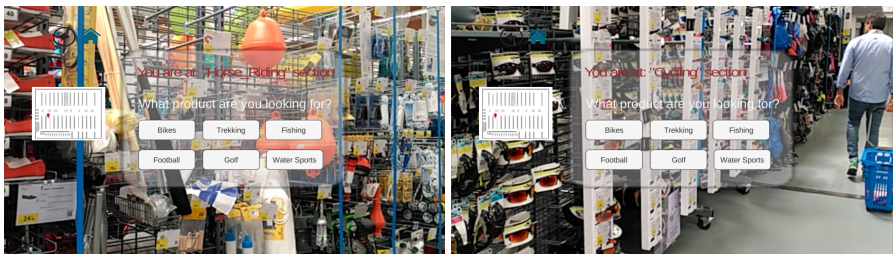

Fig. 10: Failure cases in a real scenario.

6 Conclusions

In this work, we present a mobile application using AR to improve the shopping experience in large retail stores. The contributions of this work are manifold.

First, we have combined DL techniques to locate the customer in the store. Thus, we are able to guide the customer towards the place where the desired product is located. Second, using AR, the customer is able to visualize the products and models in the store. The customer is not only able to visualize a product but also to get useful information related to the product, specifications, other available models, sizes and so on.

For future works, we aim to provide a web application where a given store can upload images of their store and the system will be able to reconfigure itself to work in that new environment. Moreover, we plan to study different compression algorithms for efficient storage and transmission of 3D models using mobile networks.

Acknowledgements This work has been supported by the Spanish Government TIN2016-76515-R Grant, supported with Feder funds. It has also been supported by the University of Alicante project GRE16-19.

References

Ronald Azuma, Yohan Baillot, Reinhold Behringer, Steven Feiner, Simon Julier, and Blair MacIntyre. Recent advances in augmented reality. *IEEE Comput. Graph. Appl.*

Olivier Pauly, Benoit Diotte, Pascal Fallavollita, Simon Weidert, Ekkehard Euler, and Nassir Navab. Machine learning-based augmented reality for improved surgical scene understanding. *Computerized Medical Imaging and Graphics*, 41(Supplement C):55 – 60, 2015. ISSN 0895-6111. Machine Learning in Medical Imaging.

Junho Ahn, James Williamson, Mike Gartrell, Richard Han, Qin Lv, and Shivakant Mishra. Supporting healthy grocery shopping via mobile augmented reality. *ACM Trans. Multimedia Comput. Commun. Appl.*, 12(1s):16:1–16:24, October 2015. ISSN 1551-6857. doi: 10.1145/2808207.

Max Ortiz-Catalan, Rannveig A Gumundsdttir, Morten B Kristoffersen, Alejandra Zepeda-Echavarria, Kerstin Caine-Winterberger, Katarzyna Kulbacka-Ortiz, Cathrine Widehammar, Karin Eriksson, Anita Stockselius, Christina Ragn, Zdenka Pihlar, Helena Burger, and Liselotte Hermansson. Phantom motor execution facilitated by machine learning and augmented reality as treatment for phantom limb pain: a single group, clinical trial in patients with chronic intractable phantom limb pain. *The Lancet*, 388 (10062):2885 – 2894, 2016. ISSN 0140-6736.

Arieh Goldman. The transfer of retail formats into developing economies: The example of china. *Journal of Retailing*, 77(2):221 – 242, 2001. ISSN 0022-4359. doi: https://doi.org/10.1016/S0022-4359(01)00044-6. URL `http://www.sciencedirect.com/science/article/pii/S0022435901000446`.

Katherine N. Lemon and Peter C. Verhoef. Understanding customer experience throughout the customer journey. *Journal of Marketing*, 80(6):69–96, 2016.

Marta Blazquez. Fashion shopping in multichannel retail: The role of technology in enhancing the customer experience. *International Journal of Electronic Commerce*, 18(4):97–116, 2014.

Kim Willems, Annelien Smolders, Malaika Brengman, Kris Luyten, and Johannes Schning. The path-to-purchase is paved with digital opportunities: An inventory of shopper-oriented retail technologies. *Technological Forecasting and Social Change*, 124(Supplement C):228 – 242, 2017. ISSN 0040-1625.

Dhruv Grewal, Anne L. Roggeveen, and Jens Nordflt. The future of retailing. *Journal of Retailing*, 93(1):1 – 6, 2017. ISSN 0022-4359. doi: https://doi.org/10.1016/j.jretai.2016.12.008. URL `http://www.sciencedirect.com/science/article/pii/S0022435916300872`. The Future of Retailing.

Silvia Likavec, Francesco Osborne, and Federica Cena. Property-based semantic similarity and relatedness for improving recommendation accuracy and diversity. *International Journal on Semantic Web and Information Systems*, 11(4):1–40, 2008.

Julie Carmigniani, Borko Furht, Marco Anisetti, Paolo Ceravolo, Ernesto Damiani, and Misa Ivkovic. Augmented reality technologies, systems and applications. *Multimedia Tools and Applications*, 51(1):341–377, 2011.

Jari Veijalainen. Mobile ontologies: Concept, development, usage, and business potential. *International Journal on Semantic Web and Information Systems*, 4:20–34, 2008.

D. Chatzopoulos, C. Bermejo, Z. Huang, and P. Hui. Mobile augmented reality survey: From where we are to where we go. *IEEE Access*, 5:6917–6950, 2017. ISSN 2169-3536. doi: 10.1109/ACCESS.2017.2698164.

Wolfgang Narzt, Gustav Pomberger, Alois Ferscha, Dieter Kolb, Reiner Müller, Jan Wieghardt, Horst Hörtner, and Christopher Lindinger. Augmented reality navigation systems. *Universal Access in the Information Society*, 4 (3):177–187, 2006.

Oded Maron and Aparna Lakshmi Ratan. Multiple-instance learning for natural scene classification. In *ICML*, volume 98, pages 341–349. Citeseer, 1998.

J. Martínez-Gómez, A. Fernández-Caballero, I. García-Varea, L. Rodríguez, and C. Romero-González. A taxonomy of vision systems for ground mobile robots. *Int J Adv Robot Syst*, 11:1–11, 2014.

Jianxin Wu, Henrik Christensen, James M Rehg, et al. Visual place categorization: Problem, dataset, and algorithm. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 4763–4770. IEEE, 2009.

G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

Jesus Martinez-Gomez, Ismael Garcia-Varea, Miguel Cazorla, and Vicente Morell. Vidrilo: The visual and depth robot indoor localization with objects information dataset. *The International Journal of Robotics Research*, 34(14):1681–1687, 2015. doi: 10.1177/0278364915596058.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2597-0. doi: 10.1109/CVPR.2006.68. URL http://dx.doi.org/10.1109/CVPR.2006.68.

Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 253–256. IEEE, 2010.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, AlexanderC. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2015. ISSN 0920-5691. doi: 10.1007/s11263-015-0816-y. URL http://dx.doi.org/10.1007/s11263-015-0816-y.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Jose Carlos Rangel, Miguel Cazorla, Ismael Garcia-Varea, Jesus Martinez-Gomez, Elisa Fromont, and Marc Sebban. Scene Classification from Semantic Labeling. *Advanced Robotics*, 30(11–12):758–769, 2016. doi: 10.1080/01691864.2016.1164621.

O. Akgul, H. I. Penekli, and Y. Genc. Applying deep learning in augmented reality tracking. In *2016 12th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, pages 47–54, Nov 2016. doi: 10.1109/SITIS.2016.17.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *Trans. Neur. Netw.*

R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training Recurrent Neural Networks. *ArXiv e-prints*, November 2012.

Zhizhong Li and Derek Hoiem. *Learning Without Forgetting*, pages 614–629. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46493-0.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Georg Klein and David Murray. Parallel tracking and mapping on a camera phone. In *Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality*, ISMAR '09, pages 83–86, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-1-4244-5390-0. doi: 10.1109/ISMAR.2009.5336495. URL http://dx.doi.org/10.1109/ISMAR.2009.5336495.

Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. Dtam: Dense tracking and mapping in real-time. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2320–2327, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-1-4577-1101-5. doi: 10.1109/ICCV.2011.6126513. URL `http://dx.doi.org/10.1109/ICCV.2011.6126513`.