Multimedia Interfaces for People Visually Impaired

Alexiei Dingli and Isaac Mercieca

Abstract In our society, there is a substantial number of visually impaired individuals. However many social mechanisms are not designed with these people in mind thus making the development of electronic assistive tools essential in order to perform basic day-to-day activities. Due to the penetration of capabilities of mobile devices, such devices have become an ideal candidate for designing solutions to aid the visually impaired. The objective of this research is to develop a multimedia user interface whose scope is to aid the visually challenged. We propose and design a product recognition system utilizing computer vision and machine learning techniques. Our system allows visually impaired individuals to identify products in grocery stores and supermarkets without any additional assistance, thus encouraging them to perform daily activities without requiring any additional help thus further promoting their independence within society. Our approach is composed of two main modules one capable of classifying grocery products using an unsupervised feature extraction methods posed by deep learning techniques while the other module is capable of recognizing products in an image using the traditionally handcrafted feature extraction algorithms. We considered multiple robust approaches to identify the one most suited for our task. Through evaluation we determined that the best approach for classification is to fine-tune a convolutional neural network pre-trained on a larger dataset. We were successful in not only surpassing our base accuracy but also obtaining an accuracy of 63 %.

Keywords Visually impaired · Computer vision · Image classification · Object recognition · Deep learning · Mobile technologies

A. Dingli (✉) · I. Mercieca
Department of Intelligent Computer Systems, Faculty of ICT, University of Malta, Msida, Malta
e-mail: alexiei.dingli@um.edu.mt

I. Mercieca
e-mail: isaac.mercieca.11@um.edu.mt

G. Di Bucchianico and P. Kercher (eds.), *Advances in Design for Inclusion*,
Advances in Intelligent Systems and Computing 500,
DOI 10.1007/978-3-319-41962-6_43

1 Introduction

Visual impairment is a crucial subject, which should be explored in depth in our visually centered world. For even the simplest of tasks, visually impaired individuals must depend on others for help. One such task is performing their daily grocery shopping, where they must compile a shopping list and entrust a supermarket employee to collect the required items for them. This not only runs the risk of such individuals being taken advantage of but also strips them of their independence. With the system we are proposing in this research, we attempt to not only solve these problems, but to also make the lives of these individuals easier by aiding them in performing such tasks completely independent of others.

Recognizing grocery items is quite a challenging task as multiple products can share similar visual attributes, one such example is a chocolate bar where one is plain and the other contains nuts while both have the same brand. Similar systems [1–4] approached this task by using a classification paradigm known as cross-dataset classification [1]. This problem paradigm refers to the case where the train and test data were not acquired from the same distribution. They used images of products, which were captured under ideal conditions for the train data, and for the test distributions they used images of the products in their real-world environment. These were referred to as in vitro [3] and in situ images [3] respectively.

These systems used a relatively small number of grocery products for example 120 [3, 4] and 30 [2] and for each one these products multiple images of the same product were used. Each class in their classification process represented a single product [2, 3]. Although multiple classification approaches were considered, the method which showed best performance was the standard bag of visual words approach [5]. However, a recent system approached our problem, by grouping multiple products depending on their particular product category one such example is pasta [1]. This system used a fine-grained classification algorithm [6] to predict the product category. Recognition was performed by matching the handcrafted SIFT [7] features. In contrast, to the other systems recognition is only performed on the products pertaining to that particular category. This system used a multi-label approach, which was not fit for our task as we attempt to identify a singular product in the individual's view. Although this system outperformed the others, the best classification accuracy achieved was 21.9 %.

By following on the method used in [1] we explored our problem by initially classifying the product and then recognizing the product by comparing the particular product with those in a precompiled database pertaining to that same product category. In our background research, we examined numerous classification approaches and observed that the recent deep learning techniques have performed quite impressively, even surpassing human performance in some cases. We approached our classification problem using convolutional neural networks (CNNs) a deep neural model which has proved to achieve a great performance for image classification tasks [8, 9]. We consider training the deep model by initializing the

weights randomly, then using learning methodologies and evaluate them by identifying the best performing approach for our task.

On the other hand, recognition was approached using Lowe's methodology [7] by matching the key points extracted from the product in the scene to those in a precompiled database. The item, which had the most number of matches, was thereby the best match for that particular product.

Our contributions in this research included a categorized dataset of in vitro images of grocery products and a methodology for identifying grocery products in a store which aids a visually challenged individual without requiring any additional assistance.

2 Aims and Objectives

The aim of this research is to aid a visually impaired individual to identify products in a grocery store. We proposed an approach where products are classified by a particular category and recognition of the product is only conducted on the training distributions pertaining to that category. To reach this aim we consider the following subsidiary aims:

We first aim to review the most recent and robust classification methodologies, commonly used in literature, in order to determine the most suitable approach for our task. The above will be achieved through the design and implementation of a sound methodology, based on approaches proposed by other researchers within the field of image classification. The basis of this methodology will consist of compiling a data collection of grocery products by acquiring image of groceries from the web and categorized these manually. Preprocessing and feature extraction techniques are then performed to attain the common attributes for classification. These attributes must be common for the products with the same category irrespective of scale, viewpoint and other transform and environmental conditions.

Secondly, we aim to investigate the most common approaches used for object recognition applications, which were examined in literature and identify the best methodology for our recognition procedure. This will be achieved through the design and implementation of an approach, commonly used throughout the literature reviewed to recognize a product in its natural environment, which in our case is most usually a shelf. The same requirements were considered for this objective as for the one highlighted above, meaning that the recognition must be invariant to most of environmental conditions.

Finally, we aim to surpass a classification accuracy of 21.9 %, which was obtained by one of the similar systems reviewed in literature [1]. In this system, the authors performed classification of grocery products using hand-crafted features and attempted to solve this particular task using a multi-label approach. This was achieved by implementing the most recent and robust image classification approaches in the design of our final system.

3 Methodology

3.1 Grocery Products Dataset

In the approach our proposed design was broken down into two components, one to predict the product's category and another to recognize the product. These required a dataset of grocery products to train our classification engine and to extract local image features for recognition. Despite the fact that in literature researchers argued that the low performance attained by their proposed system was the result of using in vitro images for the training procedure, we still used this type of images for our dataset. This is because a grocery product's packaging is continuously changed due to marketing campaigns and thereby it is nor efficient to continuously update a dataset collection of in situ images.

As we approached our classification component using deep learning methodologies this required a large collection of training images as these systems tends to attain a low performance when applied on small datasets. Though similar systems have com piled their own datasets, these were relatively small for our process. Our dataset consisted of approximately 5000 images spanning amongst 5 categories: Yogurt, Pasta, Cereal, Candy and Beverages. For each product we had a maximum of three images from three different viewpoints: front, front left angle and front right angle. These images were acquired from the web, more specifically web stores for example www.maltasupermarket.com[1] and itemMaster.[2]

3.2 Our Classification Approach

In literature, we examined that convolutional neural networks have achieved a good performance for many image classification tasks. One of the most well known models was the one proposed in [8], AlexNet [8] which led to the researchers winning the ImageNet LSVRC-2012 competition. We considered three approaches for our classification component, one where we trained our deep network model by randomly initializing the weights and the other two where we apply transfer of learning methodologies. The reason behind the latter was that these methods have achieved great performance when applied for fine-grained classification tasks [10, 11]. Although these were applied for such tasks, in literature CNNs were never utilized for cross-dataset classification more importantly trained on images in ideal conditions to be used on test data collected from a real-world environment. This module was implemented with functionality provided by the Caffe framework [12], which also provided us the ability to train our network on the GPU thereby minimizing time consumption while training. Despite the fact that our networks were

[1] http://www.maltasupermarket.com.

[2] http://www.itemmaster.com/.

trained on the GPU, this phase of our approach still consumed an enormous amount of time. The machine learning functionalities offered by scikit-learn [13] were utilized for the third approach.

3.3 Data Preprocessing

Data preprocessing was limited to only down sampling the images and subtracting the mean pixel activity. CNNs require that the images must have equal dimensionality i.e. the height must be equal to the width and vice versa. We resized the images to a size of 256 × 256 as used for AlexNet [8]. In cases where the image dimensions were not equal we added extra white pixels to the smaller dimension thereby keeping the product's structure intact. The mean pixel activity was subtracted to ensure that the network was trained on the centered raw RGB values [8].

3.4 Designing a Convolutional Neural Network

For the first approach we based our network architecture on a replica of the state-of-the-art network model, AlexNet [8]. Our network architecture consisted of 8 layers: 5 convolutional layers and 3 fully connected layers. For the first convolutional layer we used 256 filters, 128 for the second and the third, and 96 for the fourth and fifth. All convolutional layers were set a kernel of size of 3. Max pooling and local response normalization was applied to the output of the first and second layers while only mas pooling was applied to the output of the fifth convolutional layer.

We used ReLU as an activation function, as this was proven to perform well for CNNs. As examined in literature this function reduces both over fitting and time consumption while training. The first two fully connected layers were initialized with 1024 neurons while the last one only had 5 as a result of having only 5 categories in our dataset. For the sixth and seventh layer we applied dropout to minimize the issue of over fitting the network. Finally, the output of the last layer was fed to a 5-way softmax classifier.

For every iteration, we used a batch size of 64 for train distributions and a size of 35 for the test data. The batch size was dependent on our GPU memory. To train the network we used the learning procedure used in [8]. We initialized our network with a learning rate of 0.001, which decreased by a factor of 10 every 10,000 iterations. We stopped the network from training when it reached 50,000 iterations. Moreover, with every iteration we applied two data augmentation techniques, cropping with a size of 224 × 224 and horizontal mirroring to help reduce over fitting. Mirroring was only implemented for the train iterations.

3.5 Transfer of Learning Methodologies

For transfer of learning methodologies we considered fine-tuning and extracting CNN features from a pre-trained network. For both approaches we used a network pre-trained on the ImageNet ILSVRC-2012 dataset, CaffeNet (which was based on AlexNet [8]. Fine-tuning was performed by training the network on our dataset after this was pre-trained on the much larger dataset. Thus the weight would not be initialized randomly. We modified the last fully connected layer of the CaffeNet network to 5 neurons. Moreover, we initialized the learning rate multipliers for this layer to a higher value. By decreasing the global learning rate to 0.001 from 0.01 the weights for the first 7 layers adapt less quickly to our data in contrast to the last layer which had a higher learning rate multiplier. Training was performed identically to the ones in our first approach.

For the final approach we used the pre-trained network to extract a feature vector with a dimensionality of size 4096 for all the images in our train distributions. We reduced the dimensionality of these feature vectors by applying PCA dimensionality reduction. The newly reduced feature vector was fed to a LinearSVM classifier as considered for [10].

3.6 The Recognition Approach

The secondary component involved extracting SIFT [7] and SURF [14] descriptors form our dataset and store these to disk. This method was based on Lowe's approach for recognizing objects in a scene. When extracting features we applied Lowe's ratio to minimize the number of false matches. Furthermore, we used a nearest neighborhood algorithm, FLANN [15] to identify feature matches. The best match was identified by calculating the total number of matches for each comparison and identifying the product in the database, which had the most number of matches. Functionality for the feature extraction and matching was provided by the OpenCV library (Bradski).

3.7 The Client–Server System

Finally, we implemented a simple client-server system, which used the approach highlighted to recognize the grocery items. The client system was implemented with a simple interface to capture photos and sends these to the server, which in terms makes use of both components to recognize the product. The name of the product is then sent to the client, which outputs the response through a voice interface.

4 Results and Evaluation

To evaluate our proposed approach we distributed the dataset in a ratio of 75:25 for the train and validation distributions respectively. These two distributions were used for training our CNN and fine-tuned network. For classification we evaluated the performance of the approaches with the set of in vitro images in the validation set and compare these to the performance of the particular approach when applied to the in situ images. The collection of in situ images was collected manually by capturing photos of products using mobile phone. This collection was limited to a small number of popular grocery items, as we could not capture photos from local stores. The main reason for this issue was that grocery store owners did not allow us to take photos of products which also included the price. To overcome this we bought some of the products in our dataset and captured the photos while these were placed on a shelf.

By evaluating the performance of the CNN network on the validation set we concluded that the network adapted quickly to the data. One of the main reasons was that these images were too 'perfect'. In fact, we achieved good performance when evaluating the network every 10,000 iterations. However, when we tested the performance of the network on the in situ images we obtained very poor results (highest accuracy was only 34 %). In light of this we eliminated this approach from the classification module as the performance achieved on the in situ collection clearly showed that the network required to be trained on the more generic features contained in a much larger dataset and which were not available in our in vitro data.

A similar approach was considered to evaluate the performance of using a fine-tuning approach. Similarly, the evaluation on the validation set showed that the network adapted too quickly to our in vitro data. Moreover, better performance was obtained on the validation set than when using the first approach. However, the results achieved when testing the performance on the in situ data showed a much higher level of performance. In fact we achieved an accuracy of 63 %.

For the final classification approach we conducted two tests, one where we use a PCA threshold of 0.5 and another of 0.9. Similar to the previous approaches, better performance was obtained on the validation set while a lower performance was attained when applied to the validation set. Moreover, for this approach best performance was achieved using a PCA threshold of 0.9, which resulted in an accuracy of 57 % on the in situ set.

In light of the results obtained by the classification approaches considered for our problem we concluded that transfer of learning methodologies are best suited for our task. The results clearly indicate the issues that arise when training a network on a different distribution from the test set. Although both transfer of learning techniques achieved a promising performance, fine-tuning obtained the best results. Further evaluation on the results obtained using this approach indicated that worst prediction performance was obtained for the 'Yogurt' category as these products are small and with a white packaging which blends to the white background in the in vitro image. Moreover, best performance was achieved by the

'Beverage' and 'Candy' category. The reason for this is that packaging for beverages differs from the rest and the products in the 'Candy' category have a complex colored packaging, which surpasses the others. In fact most of the false predictions, were a result of predicting products as 'Candy'.

We tested our recognition component by evaluating the performance when using SIFT [7] and SURF [14] descriptors. A higher recognition rate was achieved using SIFT [7] descriptors. In fact when we used multiple viewpoints of the same product instead of using a single image, we achieved a higher recognition rate. In contrast to similar systems our product collection had products, which had very similar packaging. One such example is a yellow M&Ms and a blue M&Ms. Our results indicate that although SIFT [7] descriptors are robust, our approach could not distinguish between the two. In fact as a result of this drawback our best recognition was of 41.38 %.

We conducted one final test where we presented our system to visually challenged individuals, who tested our system and gave us feedback using a questionnaire which was e-mailed to the participants after the sessions. To test the system we visited the visually impaired participants, where we took a small number of products to be used for performing the tests. Feedback from the gathered questionnaires clearly indicate that our system was well-liked by the participants who encouraged us to keep working on it beyond the scope of this research as it provided them with a sense of independence when performing this daily activity thereby confirming that we reached our main aim, that of designing an interface to aid a visually challenged individual.

5 Conclusion and Future Work

In this research we proposed an approach, which aids a visually challenged individual to shop for groceries from local stores. Our approach consisted of two components, a classification module which aids the secondary component, recognition by only matching the products to a specific category. In contrast to similar systems, we approach our classification task using deep learning methodologies, which proved to achieve a great performance. Our results indicate that by fine tuning a CNN on our dataset we achieved promising performance, which is mostly effected by using a training set gathered from a different distribution than the test set. Recognition showed a good performance where products shared different packaging however in the case of similar packaging, this component obtained lower performance. This indicated that future work could include updating our dataset with more product images to achieve better performance for both components. Recognition could be further improved to cater for issues that arise when matching products that share an almost identical packaging. Finally, the feedback from our participants in the user evaluation clearly indicate that our system was well-liked. One participant even pointed out using video instead of photo, which could be implemented as future work.

References

1. George, M., Floerkemeier, C.: Recognizing products: a per-exemplar multi-label image classification approach. In: Computer Vision. Springer, Berlin (2014)
2. Rivera-Rubio, J., Idrees, S., Alexiou, I., Hadjilucas, L., Bharath, A.A.: Small hand-held object recognition test (short). In Applications of Computer Vision. IEEE (2014)
3. Merler, M., Galleguillos, C., Belongie, S.: Recognizing groceries in situ using in vitro training data. In: Computer Vision and Pattern Recognition. IEEE (2007)
4. Winlock, T., Christiansen, E., Belongie, S.: Toward real-time grocery detection for the visually impaired. In: Computer Vision and Pattern Recognition Workshops. IEEE (2010)
5. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of key points. In: Workshop on statistical learning in computer vision (2004)
6. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In Computer Vision and Pattern Recognition. IEEE (2011)
7. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. (2004)
8. Krizhevsky, I.S., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances In Neural Information Processing Systems (2012)
9. Arel, I., Rose, D.C., Karnowski, T.P.: Deep machine learning-a new frontier in artificial intelligence research. In Computational Intelligence Magazine. IEEE (2010)
10. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In Computer Vision and Pattern Recognition Workshops. IEEE (2014)
11. Sunderhauf, N., McCool, C., Upcroft, B., Tristan, P.: Fine-grained plant classification using convolutional neural networks for feature extraction. In: Working notes of CLEF 2014 Conference (2014)
12. Yangqing, J., Shelhamer, E., Donahue J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: arXiv preprint (2014)
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
14. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: speeded up robust features. In Computer Vision–ECCV 2006. Springer, Berlin (2006)
15. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. VISAPP (2009)