https://doi.org/10.2352/ISSN.2470-1173.2020.8.IMAWM-269

Identification of utility images on a mobile device*

Karthick Shankar†, *Qian Lin*‡, *Jan Allebach*†
† *School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN,* ‡ *HP Labs, Palo Alto, CA*

Abstract

Mobile phones are used ubiquitously to capture all kinds of images – food, travel, friends, family, receipts, documents, grocery products and many more. Often times when looking back on photos to relive memories, we want to see images that actually represent experiences and not quick convenience photos that were taken for note-keeping and not deleted. Thus, we need to have a solution that presents only the relevant pictures without showing images of receipts, grocery products etc. – termed in general as utility images. This is in the context of a photobook which compiles and shows relevant images from the photo album of a mobile device. Further, all this has to be done on a mobile device since all the media resides there – introducing the need for our system to work on low power devices. In this paper, we present a work that can distinguish between utility and non-utility images. We also present a dataset of utility images and non-utility images with images for each category mentioned. Furthermore, we present a comparison between accuracies of popular pre-trained neural networks and show the trade-off between size and accuracy.

Introduction and Related Works

We use mobile phones to capture the world around us more than ever before today. According to an article in Business Insider, 1.2 trillion images were predicted to be taken in the year 2017 [1]. This number only grows each year with phones becoming more accessible and smartphone cameras becoming more sophisticated. A large number of these images are images that don't actually represent memories. They are convenience images of certain things of which we want to take quick note. These images can be pictures of receipts, grocery products, car parking locations, documents, presentations or screenshots of text. We term these kinds of images as ”utility images”. Often times, when scrolling through our photos app on our smartphones, looking at these utility images is undesired. Thus, we bring up the concept of a photobook app that eliminates utility images and only shows pictures of real memories and experiences upon which we want to reminisce.

This work is unique in the sense that there are no prior works aiming to accomplish classification of utility images as defined above. There have, however, been numerous works for classifying each of the items in a photo individually.

For receipt and invoice classification, [2] uses machine learning with the Naive Bayes algorithm and Optical Character Recognition. There is some initial preprocessing to standardize contrast of images, following which it is sent through an OCR to capture the text. Using this, the Naive Bayes algorithm is used to classify an image as a receipt or invoice. This pipeline will not be useful for our study since the category of utility images is not restricted to receipts and invoices.

For grocery item detection, [3] allow a visually impaired user to shop at a grocery store without additional human assistance. It uses a multiclass Naive Bayes classifier, which is trained on enhanced SURF descriptors extracted from images in the GroZi-120 dataset. The study in [4] uses color and texture information in a multi-stage process: pre-selection, fine-selection and post processing. For fine-selection, it compares a classical Bag of Words technique with a more recent Deep Neural Networks approach. Both these algorithms recognize images on a grocery store shelf, which is too restrictive for the purposes of this study.

Another key component that is critical for image recognition tasks is the image itself. Publicly-available image datasets play a key role in training and testing a model quickly and provide a baseline for benchmarking different models. There are numerous general purpose image datasets like ImageNet, MS COCO, Places and CIFAR that cover a wide range of classes, such as people, scenes, things and animals [5, 6, 7, 8]. There are also datasets for specific tasks like CompCars and Flowers that have images particular to their category for classification within that category [9, 10]. There does not appear to be a public utility image dataset that has the different categories as defined above to use to train and test different models.

Furthermore, the end solution will also have to be feasible to run on a mobile device. Thabet et. al. in [11] review challenges related to image processing on mobile devices using both serial and parallel computing approaches in several emerging application contexts. In our study however, we focus on pre-training a lightweight model that can run on mobile devices as opposed to doing on-the-fly training and optimization.

As such, the contributions of this study pertaining to its research questions are as follows:

1. Build a dataset comprising of various utility image types like receipts, grocery products, car parking locations, documents, presentations and screenshots of text by concatenating various other publicly available datasets with some original images.
2. Build a classification model to classify between utility images and non-utility images that can work effectively on mobile devices.

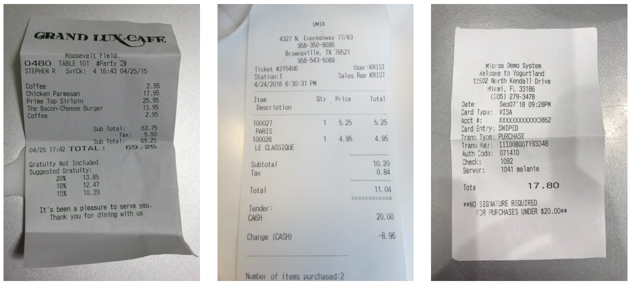
Utility Image Dataset

In this section, we give examples of images of the different categories of utility images and provide a rationale as to when these images might be taken.

Receipts

This category contains unscanned, natural images of receipts where the receipt is the main subject. A lot of publicly avail-

* Research partially supported by HP Labs, Palo Alto, CA



```
GRAND LUX CAFE

Roosevelt Field
0480  TABLE 101  #Party 3
STEPHEN R   SvrCk:  4 16:43 04/25/15

Coffee                          2.95
Chicken Parmesan               17.95
Prime Top Sirloin              25.95
The Bacon-Cheese Burger        13.95
Coffee                          2.95

               Sub Total:      63.75
                    Tax:        5.50
               Sub Total:      69.25
04/25 17:42 TOTAL:             69.25

Gratuity Not Included
Suggested Gratuity:
       20%     13.85
       18%     12.47
       15%     10.39

   It's been a pleasure to serve you.
     Thank you for dining with us
```

(a)

```
                 UMIX
       4327 N. Expressway 77/83
             956-350-8686
        Brownsville, TX 78521
             956-543-5089
Ticket #215406              User:KRIST
Station:1              Sales Rep KRIST
4/24/2016 6:30:31 PM
---------------------------------------
Item                Qty   Price   Total
 Description
---------------------------------------
100027               1     5.25    5.25
 PARIS
100026               1     4.95    4.95
 LE CLASSIQUE
                                 ------
Subtotal                          10.20
Tax                                0.84
                                 ------
Total                             11.04
                                 ======
Tender:
CASH                              20.00

Change (CASH)                     -8.96

------------------------------

Number of items purchased:2
```

(b)

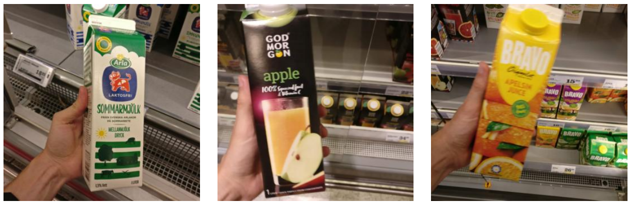
```
      Micros Demo System
     Welcome to Yogurtland
   12502 North Kendall Drive
       Miami, FL 33186
        (305) 279-3478
Date:        Sep07'18 09:28PM
Card Type:   VISA
Acct #:      XXXXXXXXXXXX3862
Card Entry:  SWIPED
Trans Type:  PURCHASE
Trans Key:   III008007193348
Auth Code:   071410
Check:       1082
Server:      1041 melanie

Tota              17.80

**NO SIGNATURE REQUIRED
   FOR PURCHASES UNDER $20.00**
```

(c)

Figure 1: Sample images of receipts.



(a) (b) (c)

Figure 2: Sample images of grocery products.

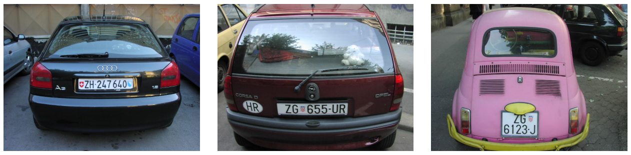
able datasets contain scanned receipts for OCR, like in [12]. The receipt dataset offered by ExpressExpense [13] however is unscanned and was thus primarily used for this part of the dataset, along with some pictures taken by us. Some samples of the images can be seen in Figure 1. Pictures of receipts can be taken for many reasons, such as expense tracking or for splitting with friends at a later date.

Grocery Products

This category contains front-on images of grocery products to keep track of the brand or product itself. There are a myriad of datasets available like [14, 15, 16, 17]. However, most of these datasets presented the items in artificial scenarios with custom platforms and lighting, or on a shelf amongst other items. Our use case is to have one primary product that faces forward so that we can gather the name of the product. Hence, the dataset in [18] was used for this part. Sample images of grocery items can be seen in Figure 2. These images can be taken when one wants to remember a particular brand of a product to buy again or when another person sends over an image of a product to let us know what it is.

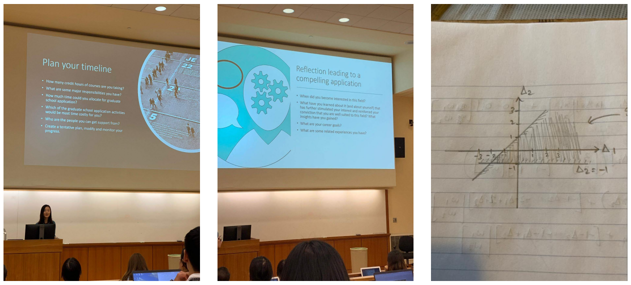
Car Parking Locations

This category contains images that show the location of where the car is parked. The images show the back of the car and a little bit of the surrounding area. License plate detection datasets work well for this case and thus, the dataset in [19] is used. These images are usually taken of rental cars in metropolitan areas where finding a car that we don't usually use can be hard. Some samples of car parking location images are shown in Figure 3. Ideally, unlike the images shown in Figure 3, the images in this category should show more of the location in which the car is situated, in order to help identify the parking location. However, even in such photos, the dominant object in the scene is a car. So the dataset in [19] should be useful for this purpose.



(a) (b) (c)

Figure 3: Sample images of car parking locations.



(a) (b) (c)

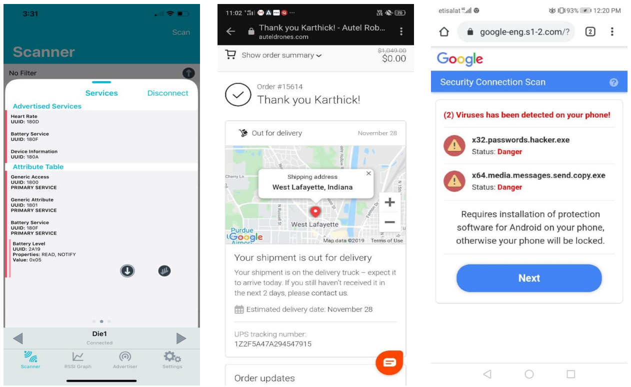
Figure 4: Sample images of documents and presentations.

Documents and Presentations

This category contains unscanned images of documents or presentations. Most datasets online contain scanned images of documents for OCR, thus making them unusable for our study. Thus, we used different smartphone cameras to capture images of documents and presentations for this category. Document and presentation images are common in the business or academic fields when one wants to keep notes of a certain page, or when a student takes pictures of presentations in class for future reference. Some samples of document and presentation images are shown in Figure 4.

Screenshots of Text

This category contains mobile phone screenshots that contain text (news articles, websites etc.). We used different smartphones to take screenshots of certain images for this category. These images are usually taken when there is a document that we want to send to somebody or there is a website of which we want to keep a note. Some samples of these images are shown in Figure 5.



3:31
Scan
Scanner
No Filter
Services Disconnect
Advertised Services
Heart Rate
UUID: 180D
Battery Service
UUID: 180F
Device Information
UUID: 180A
Attribute Table
Generic Access
UUID: 1800
PRIMARY SERVICE
Generic Attribute
UUID: 1801
PRIMARY SERVICE
Battery Service
UUID: 180F
PRIMARY SERVICE
Battery Level
UUID: 2A19
Properties: READ, NOTIFY
Value: 0x05
Die1
Connected
Scanner RSSI Graph Advertiser Settings

(a)

11:02
Thank you Karthick! - Autel Rob...
auteldrones.com
Show order summary
~~$1,049.00~~
$0.00
Order #15614
Thank you Karthick!
Out for delivery November 28
Shipping address
West Lafayette, Indiana
West Lafayette
Map data ©2019 Terms of Use
Your shipment is out for delivery
Your shipment is on the delivery truck – expect it to arrive today. If you still haven't received it in the next 2 days, please contact us.
Estimated delivery date: November 28
UPS tracking number:
1Z2F5A47A294547915
Order updates

(b)

etisalat 93% 12:20 PM
google-eng.s1-2.com/?
Google
Security Connection Scan
(2) Viruses has been detected on your phone!
x32.passwords.hacker.exe
Status: Danger
x64.media.messages.send.copy.exe
Status: Danger
Requires installation of protection software for Android on your phone, otherwise your phone will be locked.
Next

(c)

Figure 5: Sample images of screenshots.

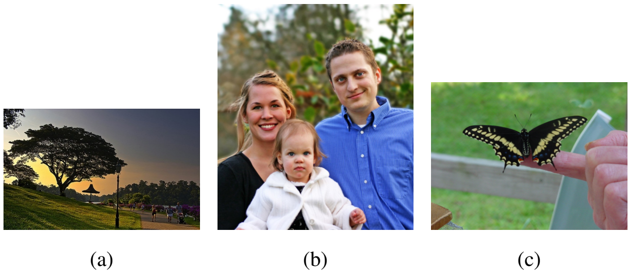


Figure 6: Sample images of non-utility images.

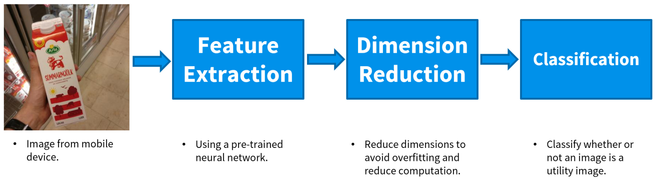

Feature Extraction
Dimension Reduction
Classification
Image from mobile device.
Using a pre-trained neural network.
Reduce dimensions to avoid overfitting and reduce computation.
Classify whether or not an image is a utility image.

Figure 7: The pipeline for image classification between utility and non-utility images.

Non-Utility Images

This category contains all the other images that don't fall into the above categories. These images represent actual memories and experiences of people. The dataset in [20] contains images that span many different categories, including people and scenery, which makes it suitable for our study. Some sample images are shown in Figure 6.

System Design

The pipeline that we use is a common one for image processing tasks and is shown in Figure 7. We first resize and crop the image as needed for the input to the pre-trained neural network. We then pass the image through the network to generate the feature vector right before the classification layer. Previous works like [21] show that the neural network features are generally separable, even if not naturally noticed by humans, like in the task of printer forensics [21]. This feature vector's dimensions are then reduced with Principal Component Analysis, followed by classification with a Support Vector Machine. Each of the steps are described in the subsections below.

Data Augmentation

Since the amount of data that we use for this model is not at the order of millions of images that is usually recommended for image processing tasks, we use simple data augmentation techniques that have roots in real world applications. The augmentations that we use are as follows:

1. Horizontal Flip
2. Vertical Flip
3. 50% Brightness
4. 150% Brightness

We use the horizontal and vertical flips since many different smartphone manufacturers chose to flip the image for better aesthetics. Similarly, the brightness level differences show the differences in exposure of different smartphone cameras.

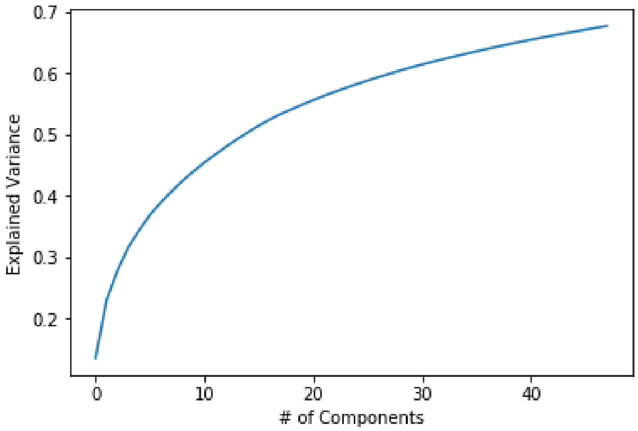


Figure 8: Explained Variance vs. Number of Principal Components

Feature Extraction

The first step of the model is feature extraction with a pre-trained neural network. One of the most popular networks to use is ResNet50 [22], with weights from ImageNet. However, as outlined in [23], ResNet50 is quite large compared to other neural networks. This extra size will make it harder to use on lower powered mobile devices. Thus, we use MobileNetV2 [24] with weights from ImageNet. According to [23], MobileNetV2 is 2× smaller in size and has 5× lower number of operations while only sacrificing 4% overall accuracy. We profile both these networks and find results which show that MobileNetV2 is still effective for our study.

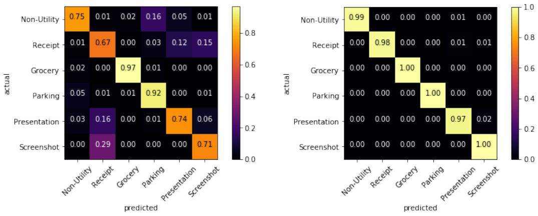
These pretrained networks have a fully-connected layer at the end that classifies a given image as one out of 1000 different classes. Since we want to extract features from these images and not classify them, we trim the network by removing the final layer. Thus, the output from the second-to-last layer is a 1×2048 vector termed as the feature vector for a given image. This vector is used as the training data for PCA as well as SVM.

Dimensionality Reduction

Since we get a large feature vector as an output from the feature extractor (pre-trained neural network), we need to reduce dimensions to ensure we avoid overfitting and reduce computation time. We use Principal Component Analysis to do this [25]. We reduce the dimensions to 48 from 2048. This number is particularly picked since it is able to explain almost 70% of the variance of the original features as shown in Figure 8. Any more would mean more risk of overfitting and any less would give poor results. Figure 9 shows the differences between PCA dimensions being reduced to 3 and 48.

Classification

For the final step, classification, we use a Support Vector Machine [26]. We show the performance of the classifier with different sets of classes - one with each utility image category as a separate class and another one with a binary classifier of utility vs. non-utility. The classifier is trained using the PCA-reduced feature vector generated through the pre-trained neural network. We use the SVM classifier that comes along with scikit-learn Python package [27].



(a) PCA-reduced dimensions = 3 (ResNet50)

(b) PCA-reduced dimensions = 48 (ResNet50)

Figure 9: Results with ResNet50 as feature extractor and SVM with 6 classes.

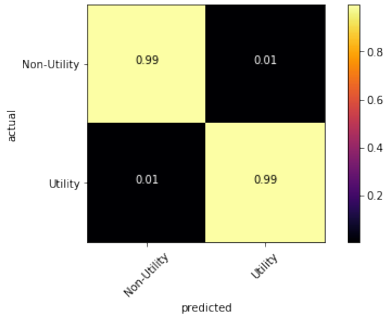


Figure 10: Confusion Matrix with ResNet50 as feature extractor and SVM with 2 classes.

Results

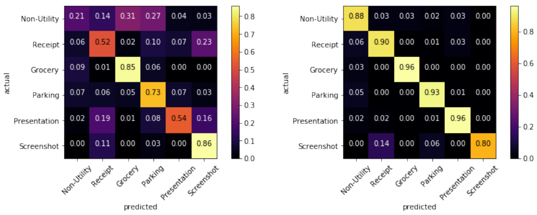
In this section we outline the different experimental results of classification with the proposed pipeline of feature extraction with a pre-trained neural network, dimensionality reduction with Principal Component Analysis and classification with a Support Vector Machine. We split this section into two – one with ResNet50 as the feature extractor and one with MobileNetV2 as the feature extractor. This is to show a theoretical best as a baseline for our comparison.

A total of 19,490 224 × 224 image segments are a part of the utility image dataset. After performing data augmentation as defined in the System Design section, we have a total of 97,450 224 × 224 image segments. This is then further split randomly into a training dataset with 80% of the images and a testing dataset with 20% of the images. The confusion matrices shown below use the testing data to test on the SVM classifier trained with the training data.

ResNet50 Feature Extractor

Figure 9 shows the confusion matrices for different PCA reduced dimensions, namely 3 and 48. Note the almost perfect classifications while using ResNet50 as the feature extractor with 48 PCA dimensions. This shows that ResNet50 performs very well when the images are general purpose images as used in this study. The performance is also better when using 48 dimensions as opposed to 3 dimensions. This is expected as 48 dimensions still carries important feature information.

If we limit classification to two classes, utility or non-utility, we see that 99% of the images are classified correctly for both classes as seen in Figure 10.



(a) PCA-reduced dimensions = 3 (MobileNetV2)

(b) PCA-reduced dimensions = 48 (MobileNetV2)

Figure 11: Results with MobileNetV2 as feature extractor and SVM with 6 classes.

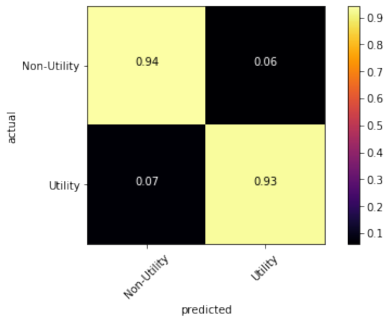


Figure 12: Confusion Matrix with MobileNetV2 as feature extractor and SVM with 2 classes.

MobileNetV2 Feature Extractor

Figure 11 shows the confusion matrices for different PCA reduced dimensions, namely 3 and 48. For 48 dimensions (Figure 11b), we see sub-par performance as compared to Figure 9b, but the lowest accuracy is only 80%, which is still acceptable. In contrast, we see a marked difference between Figure 11a and Figure 11b which clearly shows the difference between having 48 dimensions after PCA, as opposed to 3 dimensions.

If we limit classification to two classes, utility or non-utility, we see that 94% of non-utility images and 93% of utility images are classified correctly.

Conclusion

We have developed a dataset with different categories of utility images namely receipts, grocery products, car parking locations, documents, presentations and screenshots of text. The dataset also includes a set of non-utility images that represent life memories and experiences that we want to keep on our phones. We have also developed a classification model that can distinguish between utility and non-utility images for a potential photobook application that parses all the photos in the phone and only shows non-utility images. We have shown successful results for classifying each class separately, which can have Optical Character Recognition (OCR) implications for some classes, especially receipts and documents. Future works could use this to transcribe receipts and documents digitally and avoid the large space taken by images. In conclusion, we have shown that neural network features from these categories can be classified successfully.

Acknowledgments

We want to thank Kubilay Sahin and Arnav Rustagi from Purdue University for contributions to the project along the way and for suggesting improvements and further additions for future works.

References

[1] C. Cakebread, "People will take 1.2 trillion digital photos this year - thanks to smartphones," Aug 2017. [Online]. Available: https://www.businessinsider.com/12-trillion-photos-to-be-taken-in-2017-thanks-to-smartphones-chart-2017-8

[2] A. Yasser, "Classifying receipts and invoices in visma mobile scanner," *Bachelors Thesis, Linnaeus University, Vaxjo, Sweden*, 2016.

[3] T. Winlock, E. Christiansen, and S. Belongie, "Toward real-time grocery detection for the visually impaired," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, June 2010, pp. 49–56.

[4] A. Franco, D. Maltoni, and S. Papi, "Grocery product detection and recognition," *Expert Systems with Applications*, vol. 81, pp. 163 – 176, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417301227

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[6] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[7] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 487–495. [Online]. Available: http://papers.nips.cc/paper/5349-learning-deep-features-for-scene-recognition-using-places-database.pdf

[8] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[9] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," *CoRR*, vol. abs/1506.08959, 2015. [Online]. Available: http://arxiv.org/abs/1506.08959

[10] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008.

[11] R. Thabet, R. Mahmoudi, and M. H. Bedoui, "Image processing on mobile devices: An overview," in *International Image Processing, Applications and Systems Conference*, Nov 2014, pp. 1–8.

[12] J. Walter, "My personal receipts collected all over the world." [Online]. Available: https://www.kaggle.com/jenswalter/receipts

[13] ExpressExpense, "Receipt image dataset – OCR / machine learning dataset." [Online]. Available: https://expressexpense.com/blog/free-receipt-images-ocr-machine-learning-dataset/

[14] G. Varol and R. S. Kuzu, "Toward Retail Product Recognition on Grocery Shelves," *ICIVC*, 2014.

[15] P. Jund, N. Abdo, A. Eitel, and W. Burgard, "The Freiburg groceries dataset," *arXiv preprint arXiv:1611.05799*, 2016.

[16] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "RPC: A large-scale retail product checkout dataset," *arXiv preprint arXiv:1901.07249*, 2019.

[17] P. Follmann, T. Bottger, P. Hartinger, R. Konig, and M. Ulrich, "MVTec D2S: densely segmented supermarket dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 569–585.

[18] M. Klasson, C. Zhang, and H. Kjellström, "A hierarchical grocery store image dataset with visual and semantic labels," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[19] Z. K. Slobodan Ribarić, "License plate detection, recognition and automated storage." [Online]. Available: http://www.zemris.fer.hr/projects/LicensePlates/english/results.shtml

[20] EPFL, "Food image dataset." [Online]. Available: https://www.epfl.ch/labs/mmspg/downloads/food-image-datasets/

[21] K. Shankar, Z. Li, and J. Allebach, "Explaining and improving a machine-learning based printer identification system," *Electronic Imaging, Media Watermarking, Security, and Forensics 2019*, pp. 544–1–544–6.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[23] S. Bianco, R. Cadène, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *CoRR*, vol. abs/1810.00736, 2018. [Online]. Available: http://arxiv.org/abs/1810.00736

[24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017. [Online]. Available: http://arxiv.org/abs/1704.04861

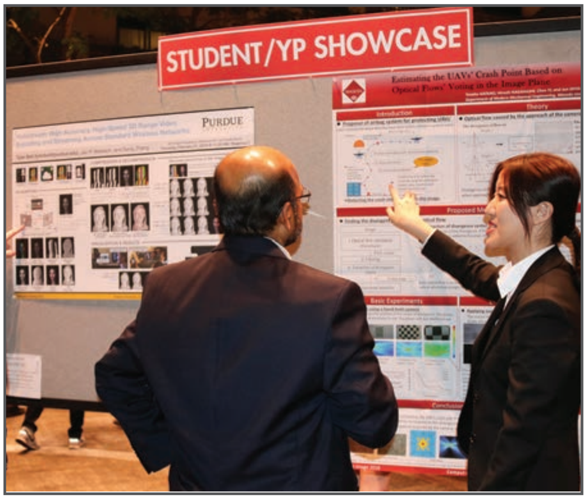
[25] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101

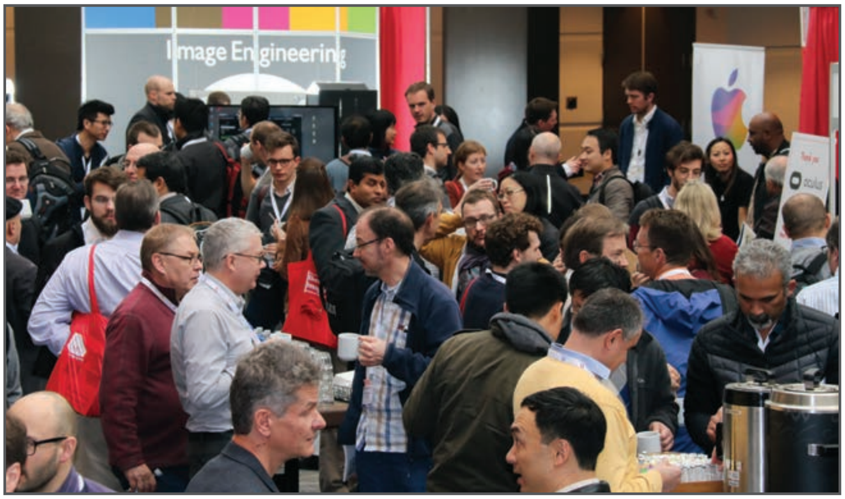
[26] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, Jun 1998. [Online]. Available: https://doi.org/10.1023/A:1009715923555

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Author Biography

Karthick Shankar is a undergraduate senior at Purdue University studying Computer Engineering. His focus is primarily in the side of distributed systems cloud engineering with an intersection in machine learning and image processing applications. He has worked on numerous projects related to this field. He has also worked with at Hulu in Seattle, WA as a Software Developer Intern. He will be pursuing further graduate studies after his final semester at Purdue. Besides academics, he was a Program Manager for Orientation Programs at Purdue and is a member of the Eta Kappa Nu honor society.





AR is a fundamental and transformative interface and platform.

20+ years from my initial working demo to today's AR displays

Optimistic that it won't take another 20 to reach ubiquitous consumer AR!