Hotel Recognition in Canada

Iryna Shvydchenko
Systems Design Engineering
University of Waterloo
Waterloo, Canada
iryna.shvydchenko@gmail.com

Spencer Hivert
Systems Design Engineering
University of Waterloo
Waterloo, Canada
spencer.hivert@gmail.com

***Abstract*—Recognizing the exact hotel from a given image can be important in human trafficking investigations. Identifying the specific location can help to further the investigation by verifying where victims have been trafficked, and where others might be trafficked in the future. This problem can be challenging mainly due to the poor image quality and the differences across rooms within the same hotel (room configurations, furniture, artwork) and the similarities between hotels. Over 1400 Canadian hotels were extracted from the Hotels-50K data set which led to over 27,000 training images coming from both travel websites and TraffickCam, a site where users are able to upload photos of their hotel rooms. Popular deep learning models including ResNet50, InceptionResNetV2, and VGG19 were adapted to this problem and were then evaluated using the top-K accuracy. An additional investigation into the brightness of the testing images was conducted, and adjusting the Gamma Correction factor of the testing images resulted in marginally improved accuracy for each of the models.**

***Index Terms*—hotel recognition, machine intelligence, human trafficking**

I. Introduction

Human trafficking has been an ever increasing problem in global organized crime [1]. According to the United Nations, human trafficking is defined as the "recruitment, transportation, transfer, harbouring or receipt of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation" [2]. In the United Nations Office on Drugs and Crime's (UNODC) global report on Trafficking in Persons, it has been shown that the most common type of human trafficking is sexual exploitation (79%), with the victims of this being mostly women and girls [3].

It is often believed that trafficking is the movement and exploitation of people across continents, however, although this is a large part of it, it has been shown that most exploitation takes place domestically [3]. These crimes are often unreported to the authorities because of the hidden nature of the crime. As this is the case, this report will focus on applying Machine Intelligence methods to help combat human trafficking within Canada. Within Canada, sexual exploitation is prevelant, especially in urban centers, with the victims including migrant women, new immigrants, at-risk youth, and those who are socially or economically challenged [4].

Furthermore, with the rapid growth of technology, traffickers have been leveraging the internet and other technologies to advertise and assist in their crimes, especially in sexual exploitation such as prostitution. In a report of how technology is used in the exploitation of domestic minor sex trafficking victims, it has been found that 63% of victims reported being advertised online through websites such as Backpage and Craigslist [5]. In these advertisements it was also found that a photo was included in the advertisement 44% of the time [5]. Many of these images are captured in hotel rooms, since they are one of the main meeting points for sexual exploitation [4]. When officials come into possession of such images, one of the main tasks is to identifying the location of the crime, which in many cases is a hotel room.

One study, Hotels-50K, has attempted to contribute to the solution of this problem on the global scale through Artificial Intelligence [6]. They curated a database of 50,000 hotels world-wide, along with 1 million training images and more than 17 thousand testing images sourced from travel websites and TraffickCam. TraffickCam is a website that allows users to upload photos of their hotel rooms to help combat human trafficking. In this report, we will use a subset of the Hotels-50k data set to attempt to identify the name and location of Canadian hotels from an input image [6].

II. Related Works

The problem of automatically recognizing hotel rooms fits within a larger effort to apply data mining, machine learning, computer vision, and natural language processing to the addressing human trafficking. In this section, we will review related efforts towards (1) Artificial Intelligence to combat human trafficking, (2) Targeted Large-Scale Image Datasets and (3) scene and place recognition.

A. Artificial Intelligence to Combat Human Trafficking

Many studies have examined the role of the Internet and related technology in facilitating human trafficking. Throughout the past decade, several studies including the work [7] explore how closely intertwined sex trafficking is with new technologies. Hughes claims that the sexual exploitation of women and children is a global human crisis that is escalated by the use of new technologies [8].

In fact, a group of experts from the Council of Europe demonstrated that the internet and sex industry are closely in-

terlinked and the volume and content of the material available online promoting human trafficking is unprecedented [9].

Early efforts leveraged data mining techniques and natural processing techniques to analyze the data on popular classified websites such as Backpage and Craigslist. These efforts hypothesized that the type of language displayed, while covert, might reveal signals of trafficking, particularly when the characteristics of the trafficking victim are used to attract clientele [9].

Other efforts have been focused on building large distributed systems to store and process available online human trafficking data in order to perform entity resolution and create ontological relations between entities [10]. Furthermore, the work in [11] uses machine learning techniques to classify advertisements for escort services by training a supervised classifier on labelled data.

These efforts focus largely on indexing online advertisements and analyzing the text and phone numbers found within or imprinted on the advertisements themselves. Moreover, private nonprofit organizations such as Thorn have built tools to scrape the online commercial sex market and have built facial recognition software to aid the victims of child sex trafficking and abuse. Thorn employs computer vision techniques which are more applicable to the hotel recognition problem but are focused on facial feature detection rather than Scene and Place Recognition.

Lastly, a study from George Washington University [6] presented a unique data set containing 50,000 hotels with over a million different images to be used for training purposes. Here, hotel recognition was framed as both a classification task (i.e., predict the label given the image) and a retrieval task (i.e., find the most similar database images to a query). As this work is directly related to the problem presented, it serves as an excellent baseline and we will work to improve their results.

B. Targeted Large-Scale Image Datasets

Historically, the computer vision community has developed datasets to support and challenge the research community. Some of the most well known datasets include ImageNet which features over 14 million images belonging to 20 thousand categories [12], MS-COCO [13] which features over 330K image and 80 object categories, and Places [14] which is a 10 million dataset for scene recognition.

These benchmarks are often used for competitions which focus on classification and retrieval methods for general categories of things which typically display high inter-class variance, meaning the categories of images are unrelated. More domain-specific datasets have been developed such as the breeds of dogs [15] and wine quality classification [16] and more applicable to our work is the Hotels-50K dataset [6].

C. Scene and Place Recognition

Scene recognition, the process of categorizing images into different bins (e.g. coast, highway, street, bedroom, and store) is a challenging problem in the field of computer vision. It is helpful to reduce the gap between computers and humans when acquiring an understanding for a scene. Thus, scene recognition has an abundance of applications in fields of computer vision and multimedia. The majority of the early approaches [17] focused on finding a mapping relating a set of low-level features (e.g., color histogram, Local Binary Pattern, and Scale-invariant Feature Transform (SIFT)) to meaningful semantic categories while more recent efforts have taken object-based, bag-of-visual-features, and attribute-based approaches [18]. Most work in this area leads to the semantic categories rather identifying the precise location of an image however this problem has received increased interest [19].

Early efforts such IM2GPS work towards estimating geographic information from a single image [20] while these efforts were later re-visited in the deep learning era [21]. These problems can be formulated as an image retrieval task where geotagged images are inferred by finding visually similar images in the dataset [6] [22].

These methods and algorithms which attempt to recognize a specific place exploit the constant geometric configurations of the landmarks typically found within these images. For instance, the Louvre in France has a very distinctive geometric shape which is more easily identifiable. Unfortunately, these geometric and matching approaches do not apply to hotel recognition. Within a hotel, the rooms may have different configurations based on the room type and furthermore, the objects found within the room may be different (i.e. artwork on the walls).

III. Dataset

A. Hotels-50K

As mentioned above, the main source of data was a data set compiled for a study in George Washington University, Hotels-50K, developed to encourage work in the hotel recognition in order to contribute to the continuous fight against human trafficking [6]. The data that was obtained contained a list of hotels, chains, a training and testing data set and a set of 'people crops' that were used in the study conducted in [6].

The training data set contained 1,027,871 training image URLs from 50,000 hotels across the world, and their hotel info. The images were sourced from two separate locations: travel websites and TraffickCam. The main website used for travel website photos was Expedia, which is a travel booking website, and the photos sourced from here are all taken professionally with great lighting and furniture decoration placement and without any clutter. On the other hand photos taken from TraffickCam, which is a website that allows users to upload images of their hotel rooms to help combat human trafficking, can be a taken with poor lighting, different furniture arrangements, perspective, clutter, etc. These images are more representative of what a trafficking investigation query photo would be like, however there are fewer images from this source. In fact, most of the training images come from the travel website (96.4%) rather than TraffickCam as seen in Figure 1.

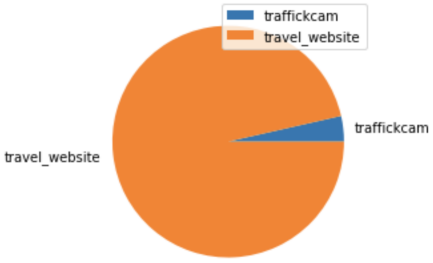

Fig. 1. Travel Website / TraffickCam Breakdown for Training Data

The difference in the two types of photos is illustrated in Figure 2, where each row shows two photos of the same hotel taken from these two different sources. It can be seen that the quality of images in the second column are much worse than the first column, with dark lighting, awkward camera angles and different decoration placement than their equivalent travel website photo.

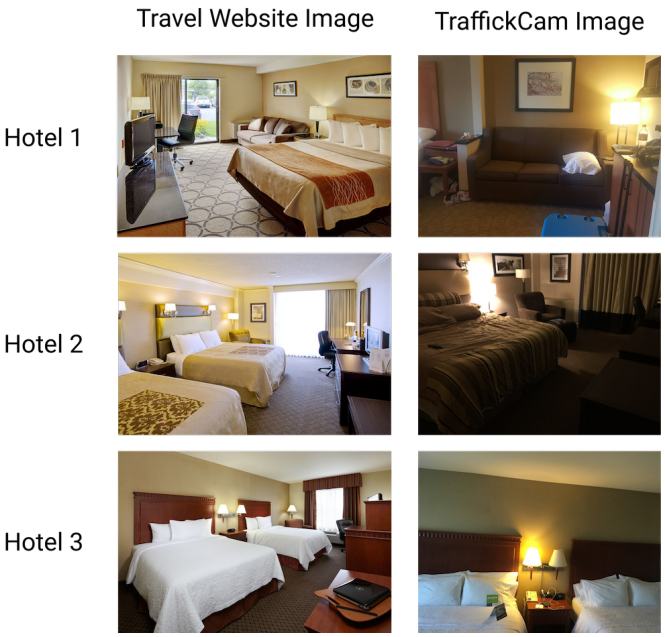

Fig. 2. Examples of Hotel Images

The testing set was very similar to the training set, with 17,153 images from a subset of 5000 of the hotels from the training set. However, all the images in the testing set provided were sourced from TraffickCam. This is done to provide a more accurate evaluation as the query images used in investigations will not be taken by professional photographers, and the TraffickCam images are much more representative of the images an investigator would use.

The data set included a list of all the hotels used in the training photos, their ID's, names, chain IDs, and latitude and longitude coordinates. Another list of hotel chains included in the data set linked the chain ID to the hotel chain it was

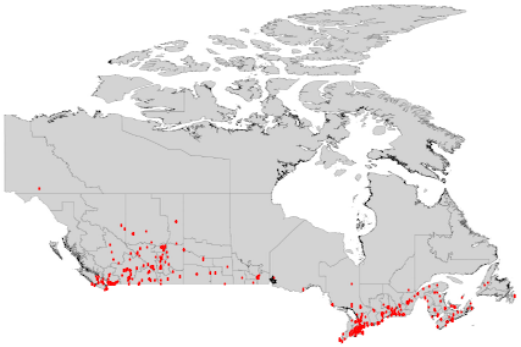

Fig. 3. Geographical Distribution of Hotels in Canada

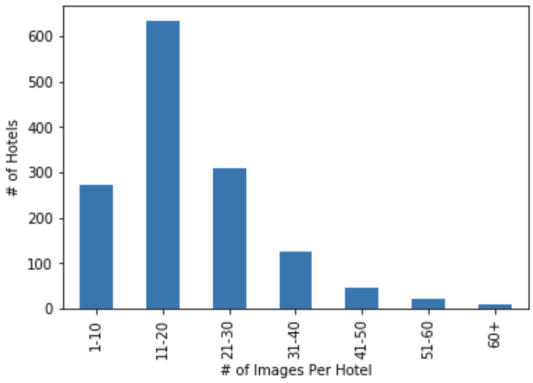

Fig. 4. Number of Images Per Hotel

from. The final component included in the data set was a set of 6485 'people crops', which are black person shaped masks that would be used to cover any human in a query image. These images were used by [6] to mask some of the training images in an attempt to get better results for more realistic query images, where it is likely that a human would be masked in a similar manner.

IV. Data Preparation

A. Data Extraction

For both the training and testing data, we began with a CSV file containing the info for each hotel, including the coordinates and the chain information for each of the hotels. The coordinates were reverse geo-coded to obtain the country and a smaller training and testing set was created with only Canadian hotels.

The decision was made to only focus on Canadian hotels due to the resources available and residence of the authors. By filtering to only include Canadian hotels, the dataset was reduced to include 1418 hotels. As seen in Figure 3, the hotels are mostly along the southern border, which is due to the larger population density [23].

As a result of the filtering, the training set was reduced to 27588 images and the the testing set was reduced to 509 images. The number of training images available for each hotel

Fig. 5. Example of Masking

ranged from only 4 images to 98 images, and the complete distribution can be seen in Figure 4.

B. Photo Masking

To replicate the real-world conditions where the regions of image containing victims are masked prior to the analysis of the image, the images can be augmented with people-shaped masks. An example of this augmentation can be seen in Figure 5. The Hotels-50k dataset provides a number of masks, which were generated using silhouettes from 'people' regions in the MS-COCO semantic labels dataset [13]. The sizes of the masks can be varied, and we have chosen three levels (low, medium, high) which correspond to the relative height of the mask. Due to time constraints, and the added complexity of the masking, it was ommitted from the experiments in this report but should be considered in future work.

V. Results

In order to set the baseline for the performance of the hotel recognition, we first define our evaluation protocol and then adapt popular convolutional neural networks to attempt to solve the problem.

A. Evaluation Protocol

The main evaluation measure used for this project was the Top-K accuracy. Top-K accuracy means that the correct class gets to be in the Top-K probabilities for it to count as correct. This was done with values of K = 1, 10, and 100. In order to measure the accuracy of each of the models, a script was run to compute the prediction array of each of the 509 images from the Canadian testing image set. The predictions that are outputted are an array of probabilities for each class.

Furthermore, a 3-fold cross validation was used for each experiment to further validate the accuracy of the results. To do this, the data was split into three groups of training and validation buckets. The validation data was selected at random form the training images for each hotel. The model was trained and validated using each of the folds independently. Then, each model was tested using the methods described above and the average and standard deviation of the results from the three models was calculated and used as the main comparison metric.

B. Experiments

The main framework used for experimentation was Kares [24] which runs on top of Tensorflow. The team also experimented with Fastai, which is built on top of PyTorch, but ultimately elected to use Keras as the K-fold cross validation was faster to implement and produced better results.

From Kares, three deep learning models were adapted with the same configuration to keep the comparison fair. A global average pooling layer was added to each model with a drop out rate of 0.7 in order to avoid over training. The training images were also augmented with many transformations such as rotation, flipping, zooming, etc. The following sections will provide an overview of the different approaches taken.

1) ResNet-50: ResNet-50 is a convolutional neural network that is trained on more than a million images from the ImageNet database [25]. ResNet first introduced the Skip Connection - extra connections between nodes in different layers of a neural network that skip one or more layers of nonlinear processing [26]. These skip connections work well for two reasons: (1) they mitigate the problem of vanishing gradient by allowing this alternate shortcut path for gradient to flow through [26] and (2) they allow the model to learn an identity function which ensures that the higher layer will perform at least as good as the lower layer [26].

2) VGG19: Similar to ResNet-50, VGG-19 is a convolutional neural network that is trained on more than a million images from the ImageNet database. The network has 19 layers and is trained to classify images into many categories. The main difference of VGG19 is that it was developed to be very uncomplicated, using a simple linear chain of layers, each with very small (3x3) convolution filters [27]. This means that less layers are needed to get similar results to competing methods [27].

3) InceptionResNetV2: The inception network was heavily engineered, and took a different approach than it's predecessors. Rather than stacking convolutional layers deeper and deeper, the network uses lot of tricks to push performance; both in terms of speed and accuracy [28].

The deeper the neural network, the more computationally expensive the convolution operations become. The inception network presented an architecture where filters with multiple size operate on the same level, essentially increasing the width of the network rather than the depth.

The inception network has seen many iterations including V2 and V3 which were also presented in [28], and the team opted to use InceptionResNetV2 which was inspired by the performance of ResNet. The hybrid network introduces residual connections that add the output of the convolution operation of the inception module, to the input. [29].

C. Experiment Summary

Table I shows a summary of the results from the experiments run, as well as the results from the Hotels-50K study [6] and what a random chance guess would be for each T op-K. However, it should be noted that the model used in the Hotel-50K study was made with the entire data set of 50,000 hotels and is thus not a fair comparison, but serves as an excellent baseline. Each result is the average of accuracies for each of three folds used in cross validation, and the standard deviation was recorded in Table II.

TABLE I
AVERAGE TOP-K ACCURACY RESULTS FOR EACH FOLD FOR EACH OF THE DIFFERENT METHODS

Hotels-50K Dataset			
(22,714 images for training and 509 images for testing)			
CNN	Top-K Accuracy (%)		
	K=1	K=10	K=100
ResNet50	7.79	20.56	48.59
InceptionResNetV2	7.07	18.01	45.78
VGG19	0.79	0.79	8.25
Hotels-50k Study[a]	8.1	17.6	34.8
Random Chance	0.0705	0.7080	7.3159

[a]The Hotels-50k results are for the global dataset.

TABLE II
TOP-K ACCURACY 3-FOLD CROSS VALIDATION SUMMARY

Hotels-50K Dataset			
CNN	Standard Deviation (σ)		
	K=1	K=10	K=100
ResNet50	1.15	1.15	3.54
InceptionResNetV2	1.29	1.67	3.75
VGG19	0.0	0.0	0.0

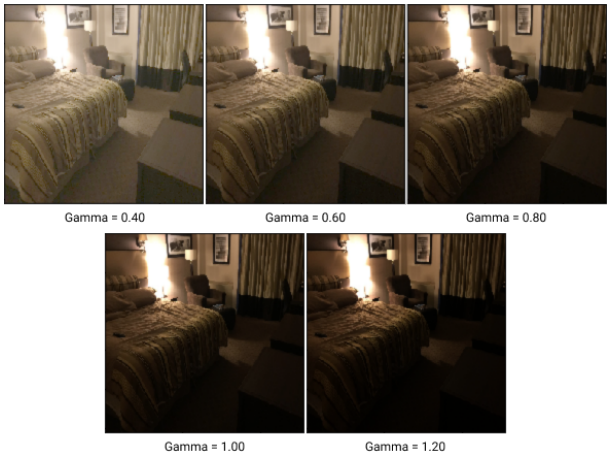

Fig. 6. Testing Image Augmentation using Gamma Correction Factor

TABLE III
COLOR AUGMENTATION RESULTS FOR K=1 ACCURACY

Hotels-50K Dataset		
CNN	Top-1 Accuracy (%)	
	No Color Augmentation	Color Augmentation
ResNet50	8.64	9.04
InceptionResNetV2	8.25	8.64

It can be seen that from the results of our experiments, that ResNet50 preformed the best, with InceptionResNetV2 a close second. VGG19 did not preform nearly as well as the other two neural networks and was only a marginal improvement over random chance. The main application of VGG is to find the object name in a given image. To do this, VGG uses small filters, which performs well for object recognition, but performs poorly with scene and place recognition.

The accuracy results for K=1 are quite poor, with the best result falling short of 10%. This poor result is no doubtly due to the larger number (1400+) of classes, the intra-class variability and inter-class similarity. Regardless, for the applications of this model, specifically for investigative purposes, it is realistic to assume that a list of potential hotels is more than adequate and as such, we will shift our focus to the top-10 and top top-100 accuracies. The results of the top-100 accuracy are very good, with ResNet50 almost having 50% accuracy. This number is much better than the reported accuracy from Hotels-50k [6] and it is likely because of the smaller number of classes. As this is the case, we recommend in the future to separate the data by counties to improve the investigation results - especially since most instances of human traffic occurs domestically and it is likely that the country of the crime will be known [3].

D. Color Augmentation

With over 1400 classes, it is unrealistic to expect these networks to have perfect hotel recognition abilities. One potential cause is the difference in the image sources used to train and test the model. As seen in Figure 2, there is a clear difference in the photos from Expedia and the photos from TraffickCam, specifically the lighting of the images.

To combat the difference in lighting in the testing images, we can augment the colour of the test photo manually in 4 different ways and let our model predict the class for each of the augmented images. A majority vote can be used to determine the final prediction of the original image. In the event of a tied vote, the prediction of the original photo will be used as the final prediction.

To augment the test images, the Gamma correction factor (which controls the overall brightness of an image) was varied. A Gamma correction factor of 1.00 represents the original image, which can be found in the second row of Figure 6. Because the hypothesis is that the testing images are too dark, there are three additional lighter augmentations and only one darker augmentation created for each image. An example of the results of these augmentations can be seen in Figure 6.

Using the voting strategy described above, the top-K for K=1 was measured for both ResNet50 and InceptionResNetV2 models. Due to poor performance of the VGG19, it was omitted from this analysis. From each method, the best performing model was chosen to participate in this experiment, and both the accuracy with and without the colour augmentation was recorded. The results can be seen in Table III, as hypothesized, increasing the brightness of the images slightly increased the accuracy of the models.

REFERENCES

[1] *RCMP national strategy to combat human trafficking*. Ottawa, Ontario Beaconsfield, Quebec: Royal Canadian Mounted Police,Canadian Electronic Library, 2012.

[2] J. Allain, "2000 Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children. Supplementing The United Nations Convention against Transnational Organized Crime," in *Slavery in International Law*, 2013, pp. 410–421.

[3] S. Chawla, K. Spangler, C. America, M. Oliveria, N. Fahmy, H. Mili, C. Asia, D. Naruka, K. Aromaa, A. Jokinen, M. Lehti, E. Ruuskanen, T. Viljanen, P. Davis, T. Leggett, and S. Malby, "Global Report on Trafficking in Persons," *Policy Analysis*, 2009.

[4] *Human trafficking in Canada*. Ottawa, Ont: Royal Canadian Mounted Police, 2010.

[5] Thorn, "A Report on the Use of Technology to Recruit, Groom, & Sell Domestic Minor Sex Trafficking Victons," *Thorn*, vol. January, 2015.

[6] A. Stylianou, H. Xuan, M. Shende, J. Brandt, R. Souvenir, and R. Pless, "Hotels-50K: A Global Hotel Recognition Dataset," 2019.

[7] D. M. Hughes *et al.*, "The demand for victims of sex trafficking," *Womens Studies Program, University of Rhode Island*, vol. 26, 2005.

[8] D. M. Hughes, "The use of new communications and information technologies for sexual exploitation of women and children," *Hastings Women's LJ*, vol. 13, p. 127, 2002.

[9] M. Latonero, "Human trafficking online: The role of social networking sites and online classifieds," *Available at SSRN 2045851*, 2011.

[10] P. Szekely, C. A. Knoblock, J. Slepicka, A. Philpot, A. Singh, C. Yin, D. Kapoor, P. Natarajan, D. Marcu, K. Knight *et al.*, "Building and using a knowledge graph to combat human trafficking," in *International Semantic Web Conference*. Springer, 2015, pp. 205–221.

[11] A. Dubrawski, K. Miller, M. Barnes, B. Boecking, and E. Kennedy, "Leveraging publicly available data to discern patterns of human-trafficking activity," *Journal of Human Trafficking*, vol. 1, no. 1, pp. 65–85, 2015.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[13] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014.

[14] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition." *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[15] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel Dataset for Fine-Grained Image Categorization," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2011.

[16] P. Cortez, J. Teixeira, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Using data mining for wine quality assessment," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009.

[17] D. L. Wang, "The time dimension for scene analysis," 2005.

[18] Y. Yuan, L. Mou, and X. Lu, "Scene Recognition by Manifold Regularized Deep Learning Architecture," *IEEE Transactions on Neural Networks and Learning Systems*, 2015.

[19] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[20] J. Hays and A. A. Efros, "IM2GPS: Estimating geographic information from a single image," in *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008.

[21] N. Vo, N. Jacobs, and J. Hays, "Revisiting IM2GPS in the Deep Learning Era," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[22] G. Baatz, O. Saurer, K. Köser, and M. Pollefeys, "Large scale visual geo-localization of images in mountainous terrain," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.

[23] J. F. Helliwell, "Measuring the width of national borders," *Review of International Economics*, vol. 10, no. 3, pp. 517–524, 2002.

[24] Chollet François, "Keras: The Python Deep Learning library," 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] A. E. Orhan and X. Pitkow, "Skip connections eliminate singularities," *arXiv preprint arXiv:1701.09175*, 2017.

[27] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[29] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.