Hotels-50K: A Global Hotel Recognition Dataset

Abby Stylianou[1], Hong Xuan[1], Maya Shende[1], Jonathan Brandt[2], Richard Souvenir[3] and Robert Pless[1]
[1]George Washington University
[2]Adobe Research
[3]Temple University
astylianou,xuanhong,mshende@gwu.edu, jbrandt@adobe.com, souvenir@temple.edu, pless@gwu.edu

Abstract

Recognizing a hotel from an image of a hotel room is important for human trafficking investigations. Images directly link victims to places and can help verify where victims have been trafficked, and where their traffickers might move them or others in the future. Recognizing the hotel from images is challenging because of low image quality, uncommon camera perspectives, large occlusions (often the victim), and the similarity of objects (e.g., furniture, art, bedding) across different hotel rooms. To support efforts towards this hotel recognition task, we have curated a dataset of over 1 million annotated hotel room images from 50,000 hotels. These images include professionally captured photographs from travel websites and crowd-sourced images from a mobile application, which are more similar to the types of images analyzed in real-world investigations. We present a baseline approach based on a standard network architecture and a collection of data-augmentation approaches tuned to this problem domain.

Introduction

In recent years, the number of images of victims of human trafficking available online has grown at an alarming rate (Bouché 2015; NCMEC 2014). Whether used for advertising or exchanged among criminal networks, these photographs can serve as visual evidence of where the victim was trafficked. Such images are often captured in hotel rooms. Identifying the hotels in these photographs to understand where a victim was (Figure 1), gives insight into trafficking operations, which is a a top priority for law enforcement (DOJ 2017).

Figure 2 shows a few example of law enforcement queries. Often the region of the images containing the victim is masked for privacy and legal reasons. Algorithms for recognition in this context must be robust to large occlusions, varying lighting conditions, and the unique perspectives of a hotel room.

This paper introduces the Hotels-50K dataset, which includes over 1 million images from 50,000 hotels around the world, designed to support efforts that address this challenging recognition task. Hotels-50K includes both professional photographs from travel websites and crowd-sourced images from a mobile application, which are more similar to the types of images analyzed in real-world investigations.

Figure 1: The Hotels-50K dataset supports the development of hotel recognition algorithms to help in investigations of human trafficking by identifying the hotel where a picture was taken.

This domain poses unique challenges compared to generic scene and place recognition tasks. These recognition problems can be grouped based on the specificity of the categories (Grauman and Leibe 2011):

1. Basic-level categories (e.g., 'building')
2. Specialized categories (e.g., 'church')
3. Exact instances (e.g., 'the Notre-Dame')

The second task ("What type of building is this?") is often referred to as *scene recognition* and the third task ("What specific church is this?") as *place recognition*. Scene recognition requires learning the shared properties of the examples in the specialized class, while place recognition requires learning the specific components and their configuration that correspond to a particular instance. Hotel recognition does not fit neatly into either task. It requires learning both the general, shared properties of all of the rooms in a particular hotel, such as its decor or star rating, or commonly used color profiles, as well as recognizing duplicated instances of furniture, art and bedding that may be used in different configurations throughout the hotel.

This paper has three main contributions. First, we propose and formulate the problem of hotel instance recognition. Second, we curate and share a data set and evaluation protocol for this problem at a scale that is relevant to international efforts to address trafficking. Third, we describe

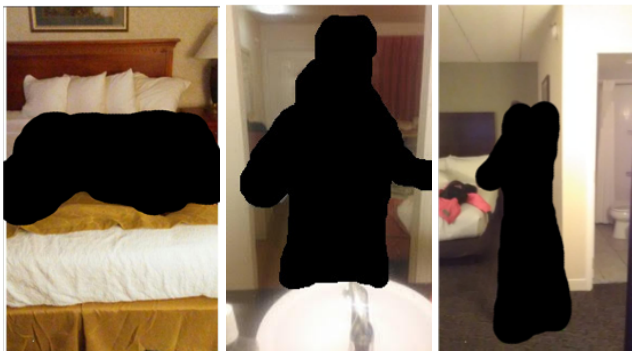


Figure 2: Example images from hotel rooms used in human trafficking investigations with the region containing the victim masked off.

and test algorithms that include the data augmentation steps necessary to attack this problem as a reasonable baseline for comparisons.

Related Work

Hotels-50k is a large-scale dataset designed to support research in hotel recognition for images with the long term goal of supporting robust applications to aid in criminal investigations. In this section, we review related efforts towards (1) AI to combat human trafficking, (2) targeted large-scale image datasets, and (3) scene and place recognition.

AI to Combat Human Trafficking. The Hotels-50K dataset and the problem of automatically recognizing hotel rooms fits within a larger set of efforts to apply machine learning, computer vision, and natural language processing to the domain of addressing human trafficking. These efforts largely focused on indexing online escort advertisements, based on locations and phone numbers in the advertisement text or imprinted on advertising images (Alvari, Shakarian, and Snyder 2017; Dubrawski et al. 2015; Kejriwal and Szekely 2017; Szekely et al. 2015). Additionally, there are larger-scale projects, such as Thorn[1] that implement approaches including facial identification for identifying victims of child sex trafficking and sexual abuse.

Targeted Large-Scale Image Datasets The computer vision community has a long tradition of developing datasets to support and challenge the research community. Some of most well-known datasets include ImageNet (Deng et al. 2009), Places (Zhou et al. 2018), and CIFAR-100 (Krizhevsky and Hinton 2009). These benchmarks drive competitions for comparing classification and retrieval methods, but because they tend to focus on general (unrelated) categories of images there have been additional efforts towards curating domain-specific datasets, including datasets of classes of cars (Krause et al. 2013) and birds (Wah et al. 2011). Most closely related to Hotels-50K

[1] https://www.wearethorn.org/

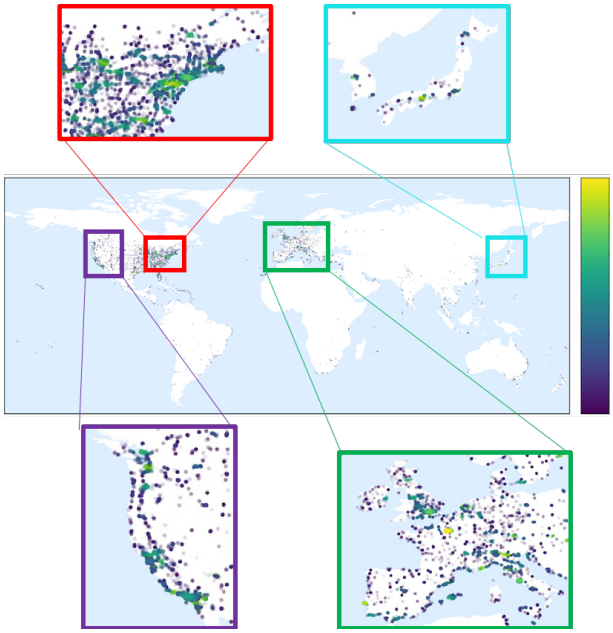


Figure 3: Geographic distribution of the Hotels-50K dataset, with a dot at every hotel location, color coded (from blue to yellow) by the local density of hotels. Images are most abundant in the United States, Western Europe and along popular coastlines.

are datasets that directly address investigative use-cases, including a database of tattoos (Ngan and Grother 2015), and a dataset of advertisements labelled by whether they include a victim of trafficking (Tong et al. 2017).

Scene and Place Recognition Recognizing the scene from which an image was captured has been a problem of great interest in the computer vision community. Most work in this area focuses on the problem of identifying the scene category (e.g., park, beach, parking lot) rather than particular locations, but recently there has been increased interest in estimating the precise geographic location of an image.

This place recognition problem can also be formulated as an image retrieval task where geotagged images serve as a database, and a query image's location is inferred by finding visually similar images in the dataset (Baatz et al. 2012; Chen et al. 2011; Crandall et al. 2009; Hays and Efros 2008; Jacobs et al. 2007; Schindler, Brown, and Szeliski 2007; Torii et al. 2013; Zamir and Shah 2010; Zheng et al. June 2009). Increasingly, methods train deep neural networks to produce similar features for images from nearby locations (Zhou et al. 2014; Arandjelović et al. 2016; Chen et al. 2017; Vo, Jacobs, and Hays 2017; Zhai et al. 2018).

Algorithms trying recognize a specific place can exploit the fact that the same objects or landmarks appear in the same geometric configuration from different viewpoints. These geometric and matching approaches do not apply to hotel recognition. Within a hotel, the rooms may have some objects that are the same (e.g., every room has the same

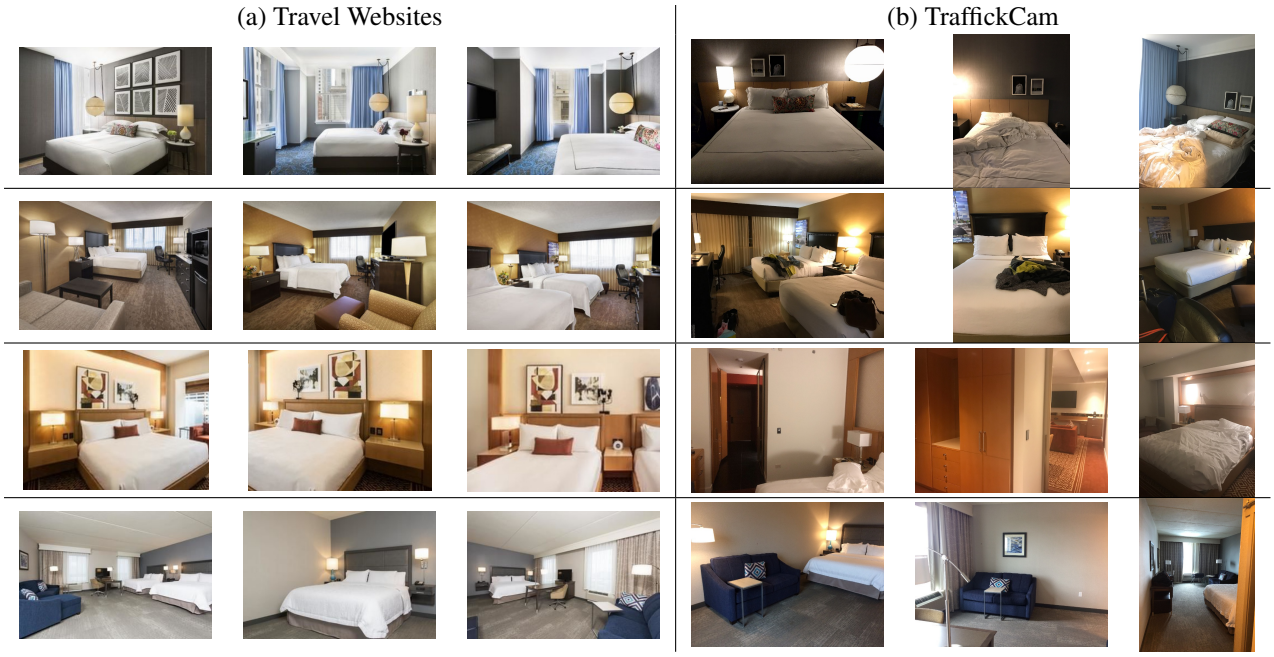


Figure 4: Comparing images across data sources shows clear differences in image quality and lighting. Each row shows images from the same hotel, with examples from (a) travel websites and (b) the TraffickCam crowd-sourcing app.

headboard), some objects that are different (e.g., different artwork on the walls), and those objects may be in different configurations from room to room (e.g., two beds vs. one or furniture on different walls).

Summary Hotels-50K follows in the tradition of large-scale datasets widely used in the computer vision and machine learning communities. This dataset will support and complement the recent trend for using AI to combat criminal activity, specifically human trafficking. The problem of hotel recognition poses unique challenges and existing methods designed for recognizing outdoor scenes or landmarks are not well-suited to the problem of discriminating between similar-looking hotel rooms.

The Hotels-50K Dataset

Hotels-50K consists of 1,027,871 images from 50,000 unique hotels around the world. Each of the images in the Hotels-50K dataset includes the following metadata: (1) hotel name (2) geographic location, and (3) hotel chain, or *Other* if the hotel property is not part of a major chain. Figure 3 shows the geographic distribution of the images in our dataset. While the dataset consists of images from around the world, the images are more densely captured in the United States, Western Europe, and coastal regions.

Data Sources The images in Hotels-50K come from two primary sources: (1) scraped from publicly available travel websites, such as Expedia and (2) captured by the crowd-sourcing mobile application, TraffickCam, which allows travelers to submit photos of their hotel room. Figure 4 shows example images from both sources captured at the same hotel. The photos from the travel websites are abundant, accounting for a majority of the images in the dataset. However, these images tend to be taken for promotional purposes, by professional photographers with excellent lighting conditions, of the nicest rooms in a hotel. These images are visually quite different from the types of images referenced in human trafficking investigations. On the other hand, while there are fewer crowdsourced images, these share more visual characteristics with the images used in real-world queries. The crowdsourced images are taken similar devices, at varying orientations, with luggage and other clutter, and without professional lighting.

Dataset Statistics Of the 50,000 hotel classes in the Hotels-50K training dataset, 13,900 have TraffickCam user-submitted images (a total of 55,061 TraffickCam images are included in the training set). There are no hotels in the dataset that have only TraffickCam images.

Figure 5 show two histograms that characterize the sampling in the dataset. Figure 5(a) shows the number of images per hotel chain for each of the 92 major hotel chains represented in the Hotels-50K dataset. Some chains have many more images than others (Holiday Inn, Hampton and Best Western), consistent with the prevalence of those hotel chains around the world. Figure 5(b) shows a histogram of the number of images per hotel broken down by the source of images (travel websites or TraffickCam mobile application). The average number of images from travel websites per hotel is 19.5. The average number of images from TraffickCam for the hotels with TraffickCam images is 4.0.

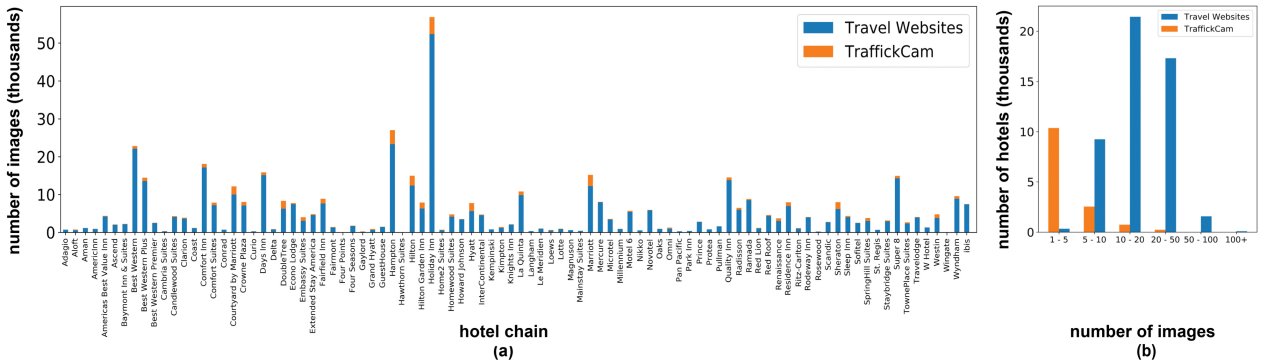


Figure 5: (a) Number of images, by source, for each of the 92 chains represented in the Hotels-50K dataset. (b) Histogram of the number of images per hotel in the Hotels-50K dataset, by the source.

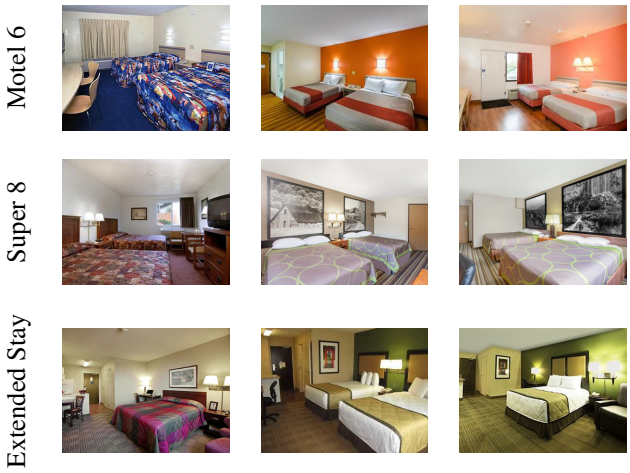


Figure 6: In each row, the first two images are from the same hotel, and the third is from a different hotel of the same chain. This highlights one of the main challenges with hotel recognition, that images within the same hotel may be visually dissimilar, while images from different hotels, especially those from the same chain, may be visually similar.

Observations While there exist discriminative patterns and unique features visible in the images from the hotels in Hotels-50K, this dataset highlights one of the main challenges in hotel recognition. There can be high intraclass variation, as not every room within a single hotel will have the same shared properties or objects – some rooms contain more amenities and some may have been renovated. On the other hand, there can be low interclass variation, especially from hotels of the same chain, making the recognition of a specific hotel difficult. Figure 6 shows a few specific examples where two rooms in the same hotel look much more different than rooms in two different hotels from the same chain.

Evaluation Protocol

Hotels-50K includes a separate test set of images to support the consistent evaluation of algorithms. Obtaining a large collection of images from real-world investigations is problematic for many reasons. However, in the images in the test set are meant to replicate, as closely as possible the type of data used in these cases.

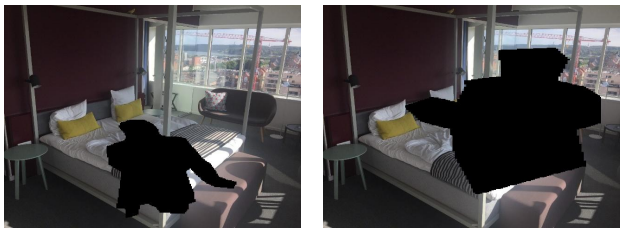


Figure 7: The images in the test set are augmented with person-shaped masks of varying size.

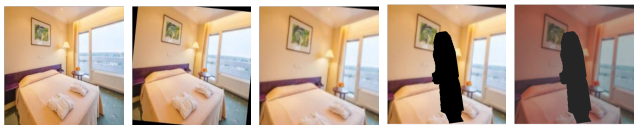
The test set consists of 17,954 images from the TraffickCam mobile application from 5,000 different hotels, which are a subset of those found in the training set. There is no overlap in the mobile app users between the training and testing sets to avoid the case of near duplicates due to multiple images from the same user with the same device captured at the same time.

To replicate real-world conditions where the regions of the image containing victims are masked prior to image analysis, the images are augmented with increasingly larger "people-shaped" masks. The masks are generated using silhouettes from 'people' regions in the MS-COCO semantic labels dataset (Lin et al. 2014). There are four levels of masking (none, low, medium high), corresponding to the relative sizes of the masked region in each image, where the largest masks can occupy up to 85% of the height of the image. Figure 7 shows examples of masked test images.

The evaluation consists of the following tasks:

Hotel Instance Recognition The goal for this task is to identify the hotel instance represented for each of the images in the test set.

Hotel Chain Recognition The goal for this task is to identify the hotel chain represented in the image. Of the test set, 13,136 images are from one of 88 major hotel chains, with the remainder in the "Other" category.



(a) (b) (c) (d) (e)

Figure 8: Data augmentation steps to better match across different lighting conditions, scales and perspectives, and in the presence of large occlusions: (a) the original image; (b) after rotation; (c) after cropping; (d) after people mask applied; (e) after color filter rendered.

Evaluation Metrics

Hotel recognition can be 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). The evaluation suite for Hotels-50K supports both variants.

For the retrieval variant, the results should be provided as a ranked list of the IDs of the 100 most similar images from the Hotels-50K dataset to each of the test images. The evaluation metric is top-K accuracy, with $K = \{1, 10, 100\}$ for hotel instance recognition and $K = \{1, 3, 5\}$ for hotel chain recognition.

For the classification variant, the results should be provided as the posterior probabilities of hotel chains or instances for each of the test images. The evaluation metrics include the average multi-class log loss (lower is better) and top-K classification accuracy with $K = \{1, 10, 100\}$ for hotel instance recognition and $K = \{1, 3, 5\}$ for hotel chain recognition.

Results

In order to set the baseline for performance on the Hotels-50K dataset, we compare two "off-the-shelf" pre-trained networks trained for object and scene recognition to a method using data and augmentation schemes specifically tailored to hotel recognition.

Models

For the pretrained models, we use the fixed feature representations and refer to these as the FIXED-OBJECT and FIXED-SCENE methods. The FIXED-OBJECT method is a Resnet-50 network trained on ImageNet (ILSVRC-2012) (He et al. 2015; Deng et al. 2009; Russakovsky et al. 2015). The feature representation is the 1001-dimensional output from the final fully connected layer. The FIXED-SCENE method uses a VGG model trained on the Places365 dataset (Zhou et al. 2018). The feature representation is the 512-dimensional output of the final pooling layer.

Our method uses the Hotels-50K training set as input to fine tune a Resnet-50 model, pre-trained for ImageNet, to output 256-D features. The training scheme is the combinatorial variant of triplet loss described in (Hermans, Beyer, and Leibe 2017).

In training, we balance the number of crowdsourced and travel website images in each batch. Additionally, we perform a set of data augmentation steps, highlighted in Figure 8. Images from the batch are randomly selected and rotated between -35 and 35 degrees, cropped between 60% and 100% of the original size, modified with color and brightness, and masked with person shaped silhouettes, similar to process used for the test data. The set of masks applied in training do not overlap with those used to generated the Hotels-50K test data and will be made available. Training parameters were selected using cross-validation. The final model was fine-tuned for 65,000 iterations with 120 images per batch.

	Instance		
	K=1	10	100
FIXED-OBJECT	0.8	0.9	1.3
FIXED-SCENE	0.2	0.8	2.4
Ours	**8.1**	**17.6**	**34.8**

	Chain		
	K=1	3	5
FIXED-OBJECT	5.0	29.0	**79.2**
FIXED-SCENE	7.2	34.2	78.7
Ours	**42.5**	**56.4**	62.8

Table 1: Retrieval results by hotel instance and by hotel chain, reported as top-K accuracy.

Retrieval

For retrieval, we compute feature representations for all of the images in the Hotels-50K training set using each method. Feature representations are also computed for each image in the test set, and the database images are ranked by cosine similarity to each test image.

Table 1 shows the image retrieval results by hotel instance and chain for all three methods. For all methods, the retrieval accuracy by hotel instance is significantly lower than the accuracy by hotel chain. This is likely due to the difficulty discriminating between particular instances of hotel chains that look similar. The chain identification task is simple enough that even the fixed methods not fine-tuned to the task achieve nearly 80% top-5 accuracy on this task. Therefore, for our remaining experiments, we focus on the more challenging problem to recognize a hotel instance.

Table 2 shows the image retrieval results for all three methods for the test images with varying sizes of image masking. Our approach has significantly higher retrieval accuracy compared to the pre-trained approaches for all tests, both with and without occlusions.

Figure 9 shows the top 5 results for several query images using FIXED-OBJECT, FIXED-SCENE and our approaches. Unlike FIXED-OBJECT and FIXED-SCENE, our model appears to encode information about the important colors and objects in a hotel room. In the top example in Figure 9, our approach finds examples from the correct hotel, as well as other images with similar blue walls and headboards. Our model also performs reasonably well even in the case where there is large amounts of clutter in the query image, as seen in the middle example in Figure 9. The last example in Figure 9 highlights the difficulty of hotel instance recognition

Occlusion:	**none**			**low**			**medium**			**high**		
	K=1	10	100	1	10	100	1	10	100	1	10	100
FIXED-OBJECT	0.8	0.9	1.3	0.3	0.4	0.7	0.0	0.0	0.0	0.0	0.1	0.4
FIXED-SCENE	0.2	0.8	2.4	0.1	0.5	1.9	0.1	0.4	1.5	0.0	0.1	1.0
Ours	**8.1**	**17.6**	**34.8**	**7.1**	**16.4**	**33.1**	**5.9**	**14.1**	**29.9**	**4.2**	**10.5**	**24.0**

Table 2: Image retrieval comparison reported as top-K accuracy.

Figure 9: The top 5 most similar results for the models trained on the Places-365 dataset, the ILSVRC dataset, and our model trained on travel website and TraffickCam images with data augmentation. Images from the correct hotel instance are highlighted in green.

Occlusion:	none	low	medium	high
FIXED-OBJECT	34.1	34.3	34.5	34.4
FIXED-SCENE	33.8	33.9	34.1	34.2
Ours	**23.8**	**24.0**	**25.4**	**27.2**

Table 3: Multi-class log loss for each method on the hotel instance classification task.

Occlusion:	none			medium		
	K=1	10	100	1	10	100
Ours -A,-I	4.7	9.6	20.0	1.8	4.0	9.4
Ours -A	**8.1**	**18.4**	**36.0**	3.5	9.2	12.8
Ours	**8.1**	17.6	34.8	**5.9**	**14.1**	**29.9**

Table 4: Ablation study reported as top-K hotel instance retrieval for our method and variants without data augmentation (-A) and without crowdsourced images (-I).

given the similarity between instances of the same hotel chain – nearly all of the top images retrieved by our model are from the correct hotel chain, but not necessarily the correct hotel.

Classification

For the classification task, we adapt the image embedding approaches used for image retrieval to report class posterior probabilities. For each method for each test image, we find the 1000 most similar images in the database using cosine similarity between the output features. The proportion of each class (hotel instance or hotel chain) in the resulting set is the estimate of the posterior probability.

Table 3 shows the multiclass log loss for each method for varying levels of occlusions in the test images. In all cases, our approach outperforms features from the pretrained models. However, there is still significant room for improved classification performance.

Ablation Study

To quantify the effects of both the inclusion of the crowdsourced data and the augmentation steps in our approach, we compare the results of variants of our method on the hotel instance retrieval task with and without significant occlusions.

This project is based in part on work supported through the National Institute of Justice (Grant 2018-75-CX-0038) and a gift from Adobe Inc.

Table 4 shows the results for the ablation experiment. We evaluate our approach without the data augmentation steps and additionally without including the crowdsourced images, which are those most similar to the real-world images. The inclusion of the crowdsourced images has a significant impact on the performance both with and without occlusions in the test image. The data augmentation steps do not have an impact on the performance in the un-occluded cases, but in the medium occlusion case, which roughly corresponds to sizes of the masked regions in real-world cases, the benefits of the data augmentation steps are apparent, increasing the top-K accuracy by more than 50% for $K = 10$.

Conclusion

In this paper, we introduced Hotels-50K, a dataset of over a million images of hotel rooms from 50,000 different hotels around the world. This dataset should further the state of the art in hotel recognition from images. We present an approach trained on the Hotels-50K dataset that outperforms fixed features from generic object and scene models. The Hotels-50K dataset, pre-trained models and code to replicate our baseline approaches can be found at `https://github.com/GWUvision/Hotels-50K`. The baseline approach is currently deployed for use by human trafficking investigators, including the National Center for Missing and Exploited Children, and novel algorithms can be quickly deployed to improve search performance in ongoing investigations.

References

Alvari, H.; Shakarian, P.; and Snyder, J. K. 2017. Semi-supervised learning for detecting human trafficking. *Security Informatics* 6(1):1.

Arandjelović, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Baatz, G.; Saurer, O.; Köser, K.; and Pollefeys, M. 2012. Large scale visual geo-localization of images in mountainous terrain. In *European Conference on Computer Vision*.

Bouché, V. 2015. A report on the use of technology to recruit, groom and sell domestic minor sex trafficking victims. Technical report, Thorn.

Chen, D. M.; Baatz, G.; Koser, K.; Tsai, S. S.; Vedantham, R.; Pylvanainen, T.; Roimela, K.; Chen, X.; Bach, J.; Pollefeys, M.; et al. 2011. City-scale landmark identification on mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, Z.; Jacobson, A.; Sünderhauf, N.; Upcroft, B.; Liu, L.; Shen, C.; Reid, I. D.; and Milford, M. 2017. Deep learning features at scale for visual place recognition. In *International Conference on Robotics and Automation*, 3223–3230. IEEE.

Crandall, D. J.; Backstrom, L.; Huttenlocher, D.; and Kleinberg, J. 2009. Mapping the world's photos. In *International World Wide Web Conference*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*.

DOJ. 2017. National strategy to combat human trafficking. `https://www.justice.gov/humantrafficking/page/file/922791/download`.

Dubrawski, A.; Miller, K.; Barnes, M.; Boecking, B.; and Kennedy, E. 2015. Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking* 1(1):65–85.

Grauman, K. L., and Leibe, B. 2011. *Visual object recognition*. Morgan and Claypool.

Hays, J., and Efros, A. A. 2008. Im2gps: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *CoRR* abs/1512.03385.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*.

Jacobs, N.; Satkin, S.; Roman, N.; Speyer, R.; and Pless, R. 2007. Geolocating static cameras. In *IEEE International Conference on Computer Vision*, 1–6.

Kejriwal, M., and Szekely, P. 2017. An investigative search engine for the human trafficking domain. In *International Semantic Web Conference*, 247–262. Springer.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

NCMEC. 2014. NCMEC amicus brief. `http://www.missingkids.com/content/dam/ncmec/en_us/documents/amicusncmecbackpage.pdf`.

Ngan, M., and Grother, P. 2015. Tattoo recognition technology-challenge (Tatt-C): An open tattoo database for developing tattoo recognition research. In *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 1–6.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252.

Schindler, G.; Brown, M.; and Szeliski, R. 2007. City-scale location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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

Tong, E.; Zadeh, A.; Jones, C.; and Morency, L.-P. 2017. Combating human trafficking with deep multimodal models. *arXiv preprint arXiv:1705.02735*.

Torii, A.; Sivic, J.; Pajdla, T.; and Okutomi, M. 2013. Visual place recognition with repetitive structures. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Vo, N.; Jacobs, N.; and Hays, J. 2017. Revisiting im2gps in the deep learning era. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2640–2649. IEEE.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology.

Zamir, A. R., and Shah, M. 2010. Accurate image localization based on google maps street view. In *European Conference on Computer Vision*.

Zhai, M.; Salem, T.; Greenwell, C.; Workman, S.; Pless, R.; and Jacobs, N. 2018. Learning geo-temporal image features. In *British Machine Vision Conference (BMVC)*.

Zheng, Y.-T.; Zhao, M.; Song, Y.; Adam, H.; Buddemeier, U.; Bissacco, A.; Brucher, F.; Chua, T.-S.; and Neven, H. June, 2009. Tour the world: building a web-scale landmark recognition engine. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhou, B.; Liu, L.; Oliva, A.; and Torralba, A. 2014. Recognizing city identity via attribute analysis of geo-tagged images. In *European Conference on Computer Vision*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2018. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(6):1452–1464.