https://doi.org/10.2352/ISSN.2470-1173.2017.10.IMAWM-165

Robust Person Recognition using CNN

Ming Chen [a], *Qian Lin* [b], *Jan P. Allebach* [a] *and Fengqing Zhu* [a]
[a] *Purdue University; West Lafayette, IN, U.S.A*
[b] *HP Labs, Palo Alto, CA, U.S.A*

Abstract

Person detection and recognition has many applications in autonomous driving, smart home and smart office applications. Knowledge about the presence of a person in the environment can be used in safety solutions such as collision avoidance, in energy conservation solutions such as turning lights and air-conditioning off when there is no person around, and in meeting and collaboration solutions such as locating a vacant room. In this paper, we present a solution that can reliably detect and recognize persons under different lighting conditions and pose based on head detection and recognition using deep learning. The system is proved to achieve good results on a challenging dataset.

Introduction

Person detection and recognition is one of the fundamental problems in image understanding. Knowing the location and identity of persons in the image leads to a lot of everyday applications. On social network and online media, detecting persons and tagging the identity has become a convenient way to share memories and organize the photos. In shopping mall and other public area, detecting and recognizing persons from surveillance cameras is an essential approach to automate applications in public safety. In a smart home and smart office environment, knowing the presence of persons can help energy harvesting, customized services, and anomaly detection.

However, detecting and recognizing persons in smart home and smart office is a challenging task. Due to the unconstrained environment and the long time span, people can have different pose, wear different clothes, and undergo various lighting conditions and occlusions.

In this paper, we investigate person detection and recognition as a whole system for smart home and smart office environment in a less constrained setting, where images contain person that have different pose, viewpoints from multiple days. We propose the use of head region instead of face or full body for person detection. We show that we can both reliably detect the person and extract powerful features that are mostly invariant to time and help with the recognition stage. To verify our person detection and recognition system, we collect ground truth annotation of head bounding box and identities for a TV series Modern Family. With the collected dataset, we show the effectiveness of our person detection and recognition system.

Related Work

Person detection and recognition, have been studied for a long time. In the bulk of previous work, person detection and recognition have been mostly treated as two separate problems, each of which has seen great progress in recent years.

Person detection is a vague term as it does not specify what body part is detected. In a lot of previous research, person detection is phrased as face detection or full body detection. Viola-Jones[1] is the textbook face detector that uses Haar feature-based cascade classifiers. This face detector is very fast, but only gives moderate detection performance. Later generic object detector based on the Deformable Parts Model (DPM)[2] has been proved to be effective for both face detection and body detection[3]. DPM models the the object by the appearance and deformation. Here, the appearance for the whole object and each part is represented by the Histogram of Oriented Gradient (HOG)[4]. The deformation calculates the deviation of parts from its ideal location relative the root. The training process will optimize the cost defined by appearance response subtracting the deformation cost at different location and scales. Recent progress in deep learning based approaches such as R-CNN[5], Fast R-CNN[6], and Faster R-CNN[7] have improved the performance of human body detection and face detection by a large margin. While it is required that frontal face, or at least a large portion of the frontal face is visible for face detection, body detection is more robust to different pose and viewpoints.

Similar to person detection, the main effort towards person recognition is on face recognition. This research field has seen great progress in the last few decades from the ones using hadcrafted features[8], to more deep learning based system such as [9]. Schroff et al.[10] use large scale proprietary data to train a network with triplet loss. Parkih et al. [11] also use triplet loss to learn an embedding of face features. These existing work mostly focus on frontal face images with little occlusion. All the above-mentioned face recognition systems more or less require face alignment as a preprocessing step.

Person detection and recognition have also been framed in the context of naming characters in TV series. The majority of this branch of work use multiple cues to recognize the persons. In [12] visual information from face and clothing appearance and textual information from the subtitles are aligned to help recognizing characters. Tapaswi et al. [13] models each episode as a Markov Random Field, integrating face and clothing appearance, speaker recognition and contextual constraints in a probabilistic model. In [13], face descriptors and multiple instance learning is applied and it is demonstrated that only using subtitles can give good results.

A closely related research area to person recognition is person re-identification[14]. In this setting, the same person is captured by cameras at different location and different time of the day. The task is to identify the person captured by one camera given a set of images captured by other cameras. It is expected that people across different time of the day and different location

*Research was supported by HP Labs, Inc., Palo Alto, CA.

wear the same clothes. Before the rise of deep learning, existing work focus on metric learning[15] and mid-level representation learning[16, 17]. Most recent work [18] have been trying to learn similarity metric through deep network using pairs of images captured from different cameras.

Recently, a dataset Person in Photo Albums (PIPA)[19] is released to help with the research in person recognition in a less constrained environment. Unlike previous research on face detection and face recognition. This work investigated the case where the frontal face is not necessarily visible. The PIPA dataset that they published contains images from every day life from thousands of persons. The author proposed an approach based on combining face recognition model and classifiers for several poselets and reported 83% accuracy for person recognition. However, achieving such a high accuracy on what seems to be a very challenging task is skeptical. A follow up paper by Oh et al. [20] investigate the flaws in the experiment protocol and found that images from the same day where people could be wearing the same clothes or even having nearly identical poses are split across the training and testing set, which explains the overly high performance. They propose to split the training and testing set according to albums or time of the day so that person recognition can be evaluated in a more realistic manner.

The datasets for face detection, face recognition, person re-identification and person recognition in general are different. The visible body parts, image quality, clothing type, and pose are all different depending on the specific tasks.

Person Detection and Recognition

A person detector is a system that generates a rectangular bounding box surrounding a person whenever a person occurs on the image. It can be applied simply for knowing the location of the person, or for presence detection where we would like to know how many persons there are in the scene. In our scenario, person detection serves as the front-end of a person recognition pipeline. In order for the following recognition engine to perform well, the detector needs to generate as many tight bounding boxes as possible, while avoiding false detections. A good detector can be crucial for the overall performance of the detection and recognition system.

However, detecting and recognizing persons in smart home and smart office is a challenging task. Due to the unconstrained environment and the long time span, people can have different pose, wear different clothes, and undergo various lighting conditions and occlusions.

In this section, we investigate person detection and recognition as a whole system for smart home and smart office environment in a less constrained setting, where images contain person that have different pose, viewpoints from multiple days. We propose the use of head region instead of face or full body for person detection. We show that we can both reliably detect the person and extract powerful features that are mostly invariant to time and help with the recognition stage. To verify our person detection and recognition system, we collect ground truth annotation of head bounding boxes and identities for a TV series Modern Family. With the collected dataset, we investigate a person detection and recognition system in smart home and smart office environment and demonstrate its effectiveness.

Training and testing results of head detection

	ZF	VGG16
Training time (hr)	10	22
Test time (ms)	59	198
mAP (%)	67.6	69.7

Head detection using Faster R-CNN

We propose the use of a less widely explored body part: head for detection and the following recognition task. By definition, head can be either frontal view, side view or even back view. In addition, most part of the head region remains unchanged from day to day as in the case of face. It also captures some contextual information like hairstyle, hair color that can help recognizing different people.

To train the head detector, we apply a state-of-the-art object detection framework Faster R-CNN[7]. Unlike previous detection framework such as R-CNN[5] and Fast R-CNN[6] that are composed of three separate stages, i.e. proposal extraction, proposal classification, and bounding box regression, Faster R-CNN is an end-to-end deep convolutional neural network that combines all three stages into a single network. The network takes image as input and predict the class of region proposals. Unlike region proposals in R-CNN and Fast R-CNN that are generated by traditional methods such as Selective Search[21], the region proposals in Faster R-CNN are generated by a branch out sub-network called Region Proposal Network (RPN). This sub-network shares the first few convolutional layers of the main detection network as described in Fast R-CNN that mostly look at edge and blob-like low-level features and can thus save a lot of computation. During training, it jointly minimizes the classification loss of the region proposals generated by RPN and the distance between the region proposals with the ground truth bounding box. We refer readers to the original Faster R-CNN paper [7] for a more detailed description of the algorithm.

The dataset we use to train the head detector is Person in Photo Albums (PIPA)[19]. All the images are crawled from Flickr and are annotated with head bounding boxes. Here we use the ground truth head bounding box annotation to train the head detector.

Table 1 shows the training time, test time, and mean average precision (mAP) of the head detector using two different pre-trained network on PIPA. We can see that with a better pre-trained network, i.e. VGG16, the detection improves by 2.1% over ZF net.

Person feature representation

As proposed in the section Head detection using Faster R-CNN, we detect head instead of face or body for the person detection task. Features in the head region will then be extracted as the person representation and fed into the recognition system. Recent progress in deep learning [22, 23, 24, 25] has shown that, the features learned from the deep network can be easily transferred to other applications that are different from the original tasks. This means that for person recognition, without the need to collect a large-scale person recognition datasets in smart home and smart office environment, we can fully utilize available public datasets. Training on these external datasets can help learn a powerful head

features that can discriminate between different identities even in smart home and smart office environment that the model has never seen.

We use the previously trained head detector to detect head regions for 10,000 identities from a recently published dataset, MS-Celeb [26], and start training the head model on the head regions. We adopt the AlexNet architecture proposed in [25]. AlexNet contains five convolution layers and three fully connected layers. Each convolution layer is followed by a max-pooling layer and ReLu layer. The first two fully-connected layers (fc6 and fc7) are of dimension 4096. Both fc6 and fc7 are followed by a dropout layer where each neuron is randomly disabled at a fixed probability in each iteration. The last fully-connected layer along with its following softmax layer has dimension of $10,000$ which is the same as the number of identities. For the cost function, we use the cross-entropy loss.

Dataset for Person Detection and Recognition System

Since there is not a public dataset that quite matches our use case, it motivates us to collect our own dataset. We purchased a popular sitcom Modern Family and label the person bounding boxes and identities for the first three of the 25-minute episodes of Series 1. The reason for choosing this TV series is that the setting is highly similar to our use case in smart home and smart office applications. Modern Family contains mostly scenes in a home environment where the characters are doing what people would do in everyday life: cooking, talking with each other, playing and so forth. The characters are not necessarily looking towards the camera, which means there is a large variety of head pose. Since there are multiple episodes, the clothing of the characters and the lighting conditions change from time to time.

There are ten main characters and a number of other people in the TV show. The main characters are of different gender, ages and hairstyle. For all but the main characters, they are treated the same as a joint class of "unknown" person. As discussed in the section Person Detection and Recognition, we suggest that head is the most effective region for person detection in an unconstrained setting. It is not only more robust to different clothing invariant than human body but also more robust to different pose and lighting conditions than face detection. Following the hypothesis, for Episode 1, Episode 2, and Episode 3, we annotated the head bounding box and the corresponding identity.

To efficiently annotate large amount of videos, we used an open source annotation tool called Vatic[27]. In Fig. 1, the annotation interface of Vatic is shown. The videos are divided into 10-second segments and the annotators work on one segment each time. The annotators are instructed to draw a bounding box around the head of a person, be it frontal view, side view, or back view, and associate it with that person's identity as long as at least half of the head is visible. Thanks to the tracking functionality integrated in Vatic, the annotator only needs to manually annotate some user-defined key frames in the video segment. The frames between consecutive key frames can be interpolated by the tool itself. When the annotation is finished, the tool samples one frame out of every 15 frames as the final output, which avoids keeping too many frames that are highly redundant.

After outputting the annotation results, we apply a simple blurry image detection algorithm based on overall edge intensity

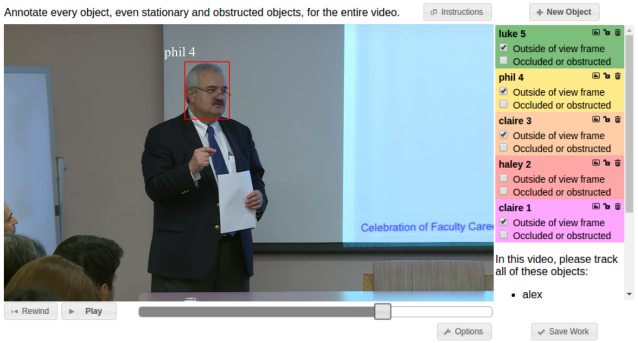


Figure 1. *The annotation interface of Vatic*

to the annotated bounding box. Blurry frames as shown in Fig. 2 are mostly due to camera and person movement. Removing these frames can help reduce ambiguity when training the recognition system.

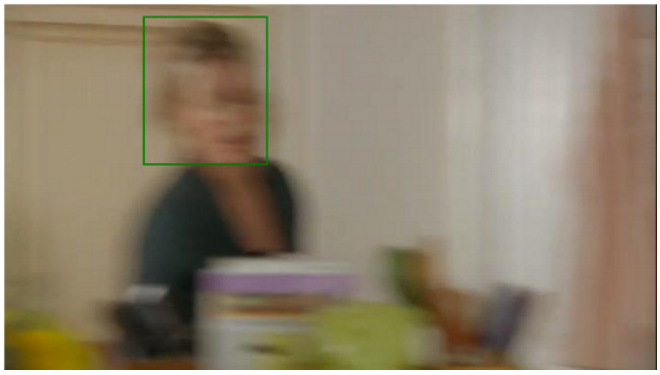


Figure 2. *Blurry image that needs to be removed*

In addition to the four labeled episodes, we also run the VGG16 head detector trained in the section Head detection using Faster R-CNN on three more episodes to create a set of unlabeled frames. This will be used for experiments on semi-supervised learning. Table 2 shows a summary of the number of annotated frames for all characters.

Person Detection and Recognition with Fully Labelled Data

In the section Person Detection and Recognition, we introduce a head detector that can reliably detect the head region of a person and a CNN model that extract features over the head. In this section, we discuss how the detector person and features are used for recognition.

In a standard supervised classification setting, there is a training set that are fully labelled. The goal is to train a classifier that fits the training set as good as possible without possibly overfitting the data. The testing set is then used to evaluate the classifier. Here, we consider the case where the system is given a set of labelled images from different identities. The task is to classify the images in the testing set.

For the training set, we assume that both the bounding box and the identities are available. The classifier is trained on fully labelld data. For the testing set, we consider both cases where

Summary of annotated data statistics

	Unknown	Alex	Cameron	Claire	Gloria	Haley	Jay	Luke	Manny	Mitch	Phil
Episode 1	316	150	351	365	455	156	503	112	199	400	432
Episode 2	50	18	244	232	196	56	174	18	29	299	239
Episode 3	483	0	456	121	163	0	342	74	168	362	496

the bounding box is available or not. When the bounding box is available, we simply extract the features using the head model and test it with the classifier. When the bounding box is not available, we run our head detector and compute the Intersection over Union (IoU) between the detected head with the ground truth head to find out the ground truth identity. If the highest matching IoU is below a threshold, which we set to 0.4 experimentally, the detection is considered as a false alarm. Otherwise, the identity corresponding to the highest matching IoU is assigned to the detected head.

The classifiers we use are Support Vector Machine (SVM) and Nearest Neighbor (NN). For SVM, we simply use the linear kernel and the following cost function:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \max\left(1 - y_i(\mathbf{w}^\mathbf{T}\mathbf{x}_i + b), 0\right), \tag{1}$$

where x_i is the 4096-dimensional head features and y_i is the ground truth label. For NN, euclidean distance is used as the distance metric.

Experiment settings

We evaluate the performance of our person detection and recognition system on the Modern Family dataset that is introduced in the section Dataset for Person Detection and Recognition System. Episode 2 is used as the testing set, while Episode 1 and Episode 3 are used as the training set. We choose to keep images from the same episodes in the same train/test subset as this will prevent images that are highly similar in the scene to appear across the training and testing set, which may lead to trivial solution as described in [20] for the PIPA dataset. For classifier training, we do not explicitly tune any parameters. The models are directly applied to the test set.

Results

We evaluate the performance of person recognition with fully labelled data. As described in the section Person Detection and Recognition with Fully Labelled Data, both cases where ground truth bounding box is available and not available for the testing set are considered. For person detection, our detector achieves precision of 85.8% and recall of 99.87%. The classification accuracy is shown in Table 3. We can see that since the recall is very high, there are only a few heads that are missing, which makes the accuracy with detected bounding box almost as high as accuracy with ground truth bounding box. Regarding the two classifiers, nearest neighbor is only around 1.5% lower than SVM, which means the head features are pretty discriminative.

We also show the two confusion matrices for the two classifiers with detected bounding box to give a little more ideas how each individual class performs. It can be seen that the confusion matrix is mostly diagonal. For both NN and SVM, the unknown class performs the worst. More specifically, unknown class for SVM is worse than unknown class for nearest neighbor. This is probably due to the fact that the unknown class is actually a mixture of different persons and is not very homogeneous. A linear SVM can easily be confused when finding the separation hyperplane for such noisy data.

The accuracy for the dataset with or without ground truth (GT) bounding box (bbox) with NN and SVM

	NN	SVM
Tested with GT bbox	86.43%	87.90%
Tested with detected bbox	86.11%	87.65%
Tested with detected bbox, evaluated among detected bbox	86.22%	87.77%

Conclusion

In this paper, we present a person detection and recognition system that can work in a barely constrained environment. We propose to use head region instead of face or body as the key body part for person detection and recognition. A head detector based on the Faster R-CNN framework is trained and can handle various pose, viewpoints, and lighting conditions. To extract rich features around the head region, we train a deep CNN model for head recognition utilizing large scale external datasets. The detection and recognition pipeline is evaluated on a challenging TV series dataset and proves to be be effective in a simple supervised learning scenario.

References

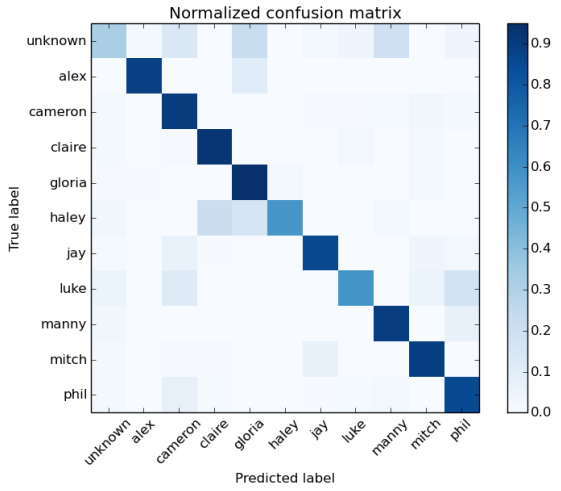
[1] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *2001 IEEE Conference on Computer Vision and Pattern Recognition*. 2001, pp. I–511, IEEE.

[2] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

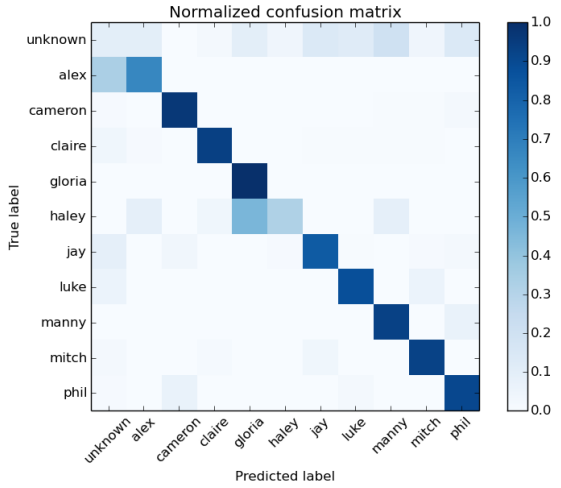
[3] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool, "Face detection without bells and whistles," in *2014 European Conference on Computer Vision*. Springer, 2014, pp. 720–735.

[4] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2005, pp. 886–893.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014,



(a) NN using detected bounding box



(b) SVM using detected bounding box

Figure 3. *Confusion matrix for face recognition with labelled data.*

pp. 580–587, IEEE.

[6] Ross Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision*. 2015, pp. 1440–1448, IEEE.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*. 2015, pp. 91–99, MIT Press.

[8] Peter N. Belhumeur, João P Hespanha, and David J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[9] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1701–1708, IEEE.

[10] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 815–823, IEEE.

[11] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference*. 2015, pp. 41.1–41.12, BMVA Press.

[12] Mark Everingham, Josef Sivic, and Andrew Zisserman, "Hello! my name is... buffy"–automatic naming of characters in tv video.," in *Proceedings of the British Machine Vision Conference*. 2006, pp. 92.1–92.10, BMVA Press.

[13] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen, "knock! knock! who is it? probabilistic person identification in tv-series," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 2658–2665, IEEE.

[14] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary, "Person re-identification by support vector ranking," in *Proceedings of the British Machine Vision Conference*. 2010, pp. 21.1–21.11, BMVA Press.

[15] Wei Li, Rui Zhao, and Xiaogang Wang, "Human reidentification with transferred metric learning," in *2012 Asian Conference on Computer Vision*. 2012, pp. 31–44, Springer.

[16] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *2013 IEEE International Conference on Computer Vision*. 2013, pp. 2528–2535, IEEE.

[17] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 152–159, IEEE.

[18] Dong Yi, Zhen Lei, and Stan Z Li, "Deep metric learning for practical person re-identification," *arXiv preprint arXiv:1407.4979*, 2014.

[19] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev, "Beyond frontal faces: Improving person recognition using multiple cues," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 4804–4813.

[20] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele, "Person recognition in personal photo collections," in *2015 IEEE International Conference on Computer Vision*. 2015, pp. 3862–3870, IEEE.

[21] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *arXiv preprint arXiv:1406.4729*, 2014.

[23] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *2014 European Conference on Computer Vision*. 2014, pp. 818–833, Springer.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. 2012, pp. 1097–1105, MIT Press.

[26] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *2016 European Conference on Computer Vision*. 2016, pp. 87–102, Springer.

[27] Carl Vondrick, Donald Patterson, and Deva Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*, vol. 101, no. 1, pp. 184–204, 2013.

Author Biography

Ming Chen received his B.Eng in Electronic and Computer Engineering from the Hong Kong University of Science and Technology (2011) and his Ph.D in Electrical and Computer Engineering from Purdue University (2016). His research interests include image processing and analysis, computer vision, and machine learning.

Dr. Qian Lin is a distinguished technologist working on computer vision and deep learning research in HP Labs. Dr. Lin joined the Hewlett-Packard Company in 1992. She received her BS from Xi'an Jiaotong University in China, her MSEE from Purdue University, and her Ph.D. in Electrical Engineering from Stanford University. Dr. Lin is inventor/co-inventor for 44 issued patents. She was awarded Fellowship by the Society of Imaging Science and Technology (IS& T) in 2012, and Outstanding Electrical Engineer by the School of Electrical and Computer Engineering of Purdue University in 2013.

Jan P. Allebach is Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. Allebach is a Fellow of the IEEE, the National Academy of Inventors, the Society for Imaging Science and Technology (IS& T), and SPIE. He was named Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IS& T, the highest award that IS& T bestows. He has received the IEEE Daniel E. Noble Award, and is a member of the National Academy of Engineering. He currently serves as an IEEE Signal Processing Society Distinguished Lecturer (2016-2017).

Fengqing Zhu is an Assistant Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, IN. Dr. Zhu received her Ph.D. in Electrical and Computer Engineering from Purdue University in 2011. Prior to joining Purdue in 2015, she was a Staff Researcher at Huawei Technologies (USA), where she received a Huawei Certification of Recognition for Core Technology Contribution in 2012. Her research interests include Image processing and analysis, video compression, computer vision and computational photography.