A Multi-Task Grocery Assist System for the Visually Impaired

Peter Zientara, Siddharth Advani, Jack Sampson, and Vijaykrishnan Narayanan

I. Introduction

ACCORDING to the World Health Organization, "285 million people are estimated to be visually impaired worldwide." [**?**] Several technologies such as automatic text readers, Braille note makers, and navigation assist canes have been developed to assist the visually impaired. Concurrent advances in computer vision and hardware technologies provides opportunities for a visual-assist system that can be used in multiple contexts.

As part of the Visual Cortex on Silicon program, we have been developing algorithms, hardware platforms and interfaces to assist the visually impaired with a focus on grocery shopping. Grocery shopping is an essential activity in our daily lives that involves various interconnected activities. These include checking the pantry for current inventory, making a shopping list based on planned meals, getting to the store, and making opportunistic and impulsive purchases in response to signage at the store. Each one of these activities poses a significant challenge without visual cues. Consequently, our ThirdEye prototype enables a combination of hardware software mechanisms to interpret these visual cues and communicate them to a visually impaired user as verbal or vibrational feedback.

A potpourri of different vision algorithms spanning brain-inspired algorithms, structured feature extraction techniques and deep learning approaches is used to support the automation of the different visual tasks. While both the origins and implementations of these algorithms are diverse, many share a common feature in that they work to reduce the potentially vast search spaces that vision problems can involve. Brain-inspired solutions, such as saliency algorithms, help to focus attention onto only specific parts of a complex image, thereby significantly reducing the computational effort to process the input and can act as a filter for subsequent steps. Other brain-inspired solutions such as GIST provide a contextual reference to the image and prime the decision making with that information, thereby filtering the model space of likely objects to be found.

In composition, these algorithms can be very powerful. Consider searching for an object in an aisle. First, steps such as saliency can help you reduce the effort spent examining the floor, empty shelves and other parts of an image that don't strongly register as potential grocery objects. Then, by understanding the context of specific grocery aisles instead of the entire store, complexity can be further pruned. For example, if milk is in your shopping list, the system can be primed to only search for milk when you arrive at the dairy aisle and to furthermore limit the number of distinct types of possible objects considered during classification to those likely to be in a dairy aisle. Such optimizations considerably reduce the computational load on the assistive vision system. In conjunction with these algorithmic advances, we have developed customized hardware solutions to make these operations more power-efficient as well as provide real-time feedback to the user.

Fig. 1. Applications and underlying technologies

The rest of this article describes the following set of visual assistive functions: (1) Identifying objects in a pantry including misplaced items; (2) Identifying other shoppers when navigating in the store or when navigating to the stores; (3) Locating packaged objects from the grocery shelf and picking them; (4) Assisting in identifying items from prepared food sections.

II. Platform and Interfaces

One of the most important aspects of any technology is how a human interacts with it. Having an intuitive, simple, and functional interface can often be the difference between a successful device and one which isn't. This importance on a interface is even more important when trying to help someone with an impairment. Not only does the interface have to be all the things mentioned above, but it also has to be robust for different environments. In the case of assisting a person with visual impairment this means being able to handle cases which people without visual impairment handle without even realizing they do. Such a scenario would be when reaching for

Authors are with the School of Electrical Engineering and Computer Science, Pennsylvania State University, PA, USA (email: {paz117, ska130, sampson, vijay}@cse.psu.edu)

Manuscript received September 16, 2016; revised Month Date, 2016; accepted Month Date, 2016.

a product if you momentarily move your head in a different direction.

Our visual assistive system consists of several interacting components:

- One of the devices that we use an off the shelf androdi powered smart glasses. The smart glasses have both a camera and a built in headset along with networking capability.
- We use a specially designed prototype glove that has been modified to have both a camera attached to it, as well set of vibration motors.
- Smart Cart. A shopping cart that can be equiped with a moderate level of computer and a variety of sensors that would be provided by the retail location.
- High-performance server machine with both GPU and FPGA integration running accelerating compute with custom algorithms and architectures.

A. Interfaces

For our assistive technology we employ two main modes of providing this feedback and guidance to the user. these modes are auditory feedback and tactile feedback. To provide this feedback to the users we use glove and glass listed above.

1) Smart Glasses: The off the shelf smart glasses provide the system with a camera in the viewpoint of the persons head as well as network connectivity and speakers for audio feedback. In the assistive system the glasses are mainly used to guide the person at the aisle level to be infront of their intended/desired product. The commands such as ”left, right, forward, back” provide the direction.

2) Custom Glove: The custom glove has both a camera and a series of vibration motors. This camera that is on the glove allows the system to have the view point of where the person is reaching. This view point may be different from that of the camera mounted on the headset and is critical to being able to provide guidance all the way to physically picking up the intended product. The vibration motors attached to the glove allow the system to provide subtle feedback to the user the convey to them which when they would have to move their hand to be able to grab the desired product. An example of this would be buzzing the right motor to indicate a rightward motion or the top motor to indicate the person needs to lift their hand.

B. Using the System

While certain aspects of system use differ across the particular tasks it supports, the modes and mechanisms employed during grocery shopping provide good coverage of typical operations, and we describe them in detail below. In the our system these the auditory feedback combined with the haptic feedback from the glove to provide the needed assistance to the shopper.

Figure 2 shows how a person would wear and interface with the asssitive system. Such a system is ideal to provide guidance and assistance in a retail type setting. ¡— please change this line.

Fig. 2. A person using an assistive system using multiple modes of feedback.

C. Challenges

Creating a truly assistive system with a variety of interfaces presents a series of challenges, not all of which are intially obvious. These challenges include, guiding the person through the store which includes, localization, obstiacle and person avoidance. Other challenges are user centric. These include adapting the frequency of guidance commands to the speed at which the person is moving, reconciling different camera views to provide correct guidance, and having enough computational power to keep the system real-time. Through various methods these challenges can be solved. To solve the problem of guiding the person through the store, the smart cart could be equipped with various sensors. This could include cameras that not only have RGB information, but even depth and possibly thermal. Also the use of localization technologies such as indoor gps, and bluetooth beacons around the store have the ability to track the user and provide the needed level of localization to the system. Another challenge that arose came from having two camera views that aren't always in alignment with one another. An example of when is occurs is when shopping and you go to grab a product, you might look away while still reaching in towards your intended product. This poses a challenge to a system giving guidance based on the view from those cameras. One possible solution to this issue would be the addition of sensors to the glasses and the glove. The addition of an IMU and Magnomter to both of the edge compute give ability to correctly give guidance in these cases. An a example how this would work is if the headset camera view indicated the person needed to move right, but the glove camera was pointing straight. The system would be able tell the user to turn just their head to align the two views. The biggest challenge that exists in a guidance system is being able to keep up with the real-time demands of the user. With an assistive system, meeting solving this is crucial. In order to do this effectively, the system as a whole must leverge every availble compute power including that available at both the edge devices, the local infastructure in addition to a powerful cloud compute platform. As stated earlier, for our cloud compute device use a high-performance server that is enabled/enhanched with both FPGAs and GPUs.

By leverageing custom architecures and exploiting parrallel algorithms we are able to process 1080p at 50fps. While this may seem like it meets the real-time constratint it doesn't. Since a server needs to be able handle multiple connections at once, just the acclerated back end would handle 50 streams at 1fps. To make up this gap in performance, tricks need to be played at the local and edge compute devices to make this difference imperceivable. Some of the compute that can be done at the front end mainly involve the filtering of data. For instance, a local intfrastructure would be able to run the images being streamed back to the server through one of the saliency algorithms expalined above. Additionally the edge device could use its sensors to only send a frame when the user has moved enough that the scene needs to be recomputed on fully. Additionally once the desired products are found, the local infastructure or edge device would be able to run a computationaly less intense tracking algorithm to be able to continue to guide the user toward the product between communications with a the cloud backend.

III. Beyond Deep Learning: Pragmatic Optimizations for Constrained Resources and Limited Time

Deep learning architectures like belief networks and neural networks are accelerating the pace of innovation in various industries such as autonomous systems, retail shopping and social media. These complex computational models are used to churn large amounts of data so as to make predictions on new observations. For example, consider an autonomous vehicle driving through a busy street. The decision of stopping the vehicle will depend upon whether there is an impeding obstacle or a red signal. More importantly this decision will be needed to made in a small amount of time and may be based upon multiple noisy sensory inputs.

Convolutional Neural Networks (CNNs) have become extremely popular and being used to solve a variety of image recognition and computer vision tasks. More recent and advanced CNN architectures have become deeper and more complex having 10 to 20 layers of Rectified Linear Units, hundreds of millions of weights, and billions of connections between units. The reader is pointed to [1] for insights on deep architectures in general and [2] for CNN-based learning and their recent advances.

While CNNs are an important thrust of research, they tend to be computationally expensive and deploying them on mobile platforms results in huge memory overheads. Another important insight recently unearthed by [3], is that the accuracies of CNNs can saturate after a few million images of training data. Also the overall efficacy of the image recognition pipeline is contingent upon having a good region proposal scheme that feeds regions of interest (RoIs) into the CNN. Considering these challenges, a variety of strategies can be used to augment the capabilities of neural networks and we outline a few below.

A. Visual Attention

Humans process and respond to only certain streams of visual information depending upon the task at hand. From a systems perspective, visual attention can be used as an efficient mechanism to prioritize data processing. Pixel-level saliency models such as Attention by Information Maximiztion (AIM) [4] can be used in automatic household pantry organization and maintenance, particularly as part of assist systems for the visually impaired. Computationally, AIM determines visual salience based on the amount of information present in local regions of the image within the context of its surrounding region. Suppose a product (say cookies) is wrongly placed in a shelf that stores products of another type (say shampoo), the segment of the shelf image containing cookies is "less likely" (higher self-information) to appear in the scene which mostly has image patches of shampoo, and therefore it is easily distinguishable or is considered "salient". This is shown in Figure 3. Once a salient region is detected, a second stage of object classification can be deployed to identify the wrongly placed object.

(a) Original image (b) Thresholded saliency map

Fig. 3. Saliency used for missplaced item detection.

Figure 5 shows another example of how we can use a combination of saliency and structured features like SURF to product intelligent segmentation of items in a grocery shelf. This is in contrast to other region proposal schemes that tend to work well with homogenous objects like people and cars but struggle to produce good intersection over unions (IoUs) over small and closely placed grocery items.

(a) Original image (b) Segmented heatmap

Fig. 4. SURF keypoint clustering used for product segmentation.

B. Redundancy

A lot of visual scenes exhibit redundancy in some form or the other. For example, in a grocery aisle there are a lot of similar looking products like cereal boxes, detergent bottles, etc. Rather than processing each of these items as independent entities, we can localize similar RoIs, run our classification engine on only one of them and then assign the corresponding label to the entire group of similar RoIs. Figure 5 illustrates this flow where AIM is used to generate initial seed RoIs that are then coupled with Speeded Up Robust Features (SURF) keypoint matching to generate a list of RoIs that are similar in structure.

C. Context

While deep learning models use learnt features to recognize objects in a scene, another contrasting approach is to use graphical models that build hierarchical representations of objects [5]. Compositional rules can be used to build context cues to recognize objects never seen before. For example, an object having four wheels can be classified as a vehicle even if it is a new model of a car that the classifier was never trained on. When it comes to recognizing objects in video streams, spatial context can play a huge role in reducing the workload on computationally intensive models such as CNNs. For example, in [6], the authors proposed a Bayesian network called Visual Co-occurrence Network (ViCoNet) where objects were represented as nodes and edges represented spatial relations between them. This network was then used to improve the performance of their system as well as the recognition rates. Figure 6 shows an example of such a graph, which could very well represent a fresh-fruit section of a grocery shop.

D. Multimodal Fusion

Humans use multisensory information from different sensory systems and combine it to influence perception, decisions, and overt behavior [7]. Wearables can be used in a similar fashion to help users in different tasks. A rich topic of exploration is figuring out a way to fuse multi-sensor information, especially data from vision that is fundamentally two-dimensional with a temporal unidimensional stream of data from other sensors to make predictions of the current state of the user. Multi-sensor information coming in from different devices can be streamed to distributed networks that can then make real-time updates. In Figure 7, we illustrate data recorded from a wearable device while two users walk in three different directions. As can be seen, these sensors are sensitive enough to be used as localization cues.

IV. Hardware Support

Even with significant investments in algorithm development and selection, the computational costs of running a visual assist system on traditional computing platforms can still be significant and a potential impediment to practical deployment. In this section we describe efforts to implement custom designs for assistive vision, leverage new, non-traditional architectures, and forecast the impact of emerging technologies that can further improve the efficiency and effectiveness of visual assist systems from the wearable front ends and mobile edge computing platforms through the cloud-hosted back-ends and databases.

A. Custom Chips

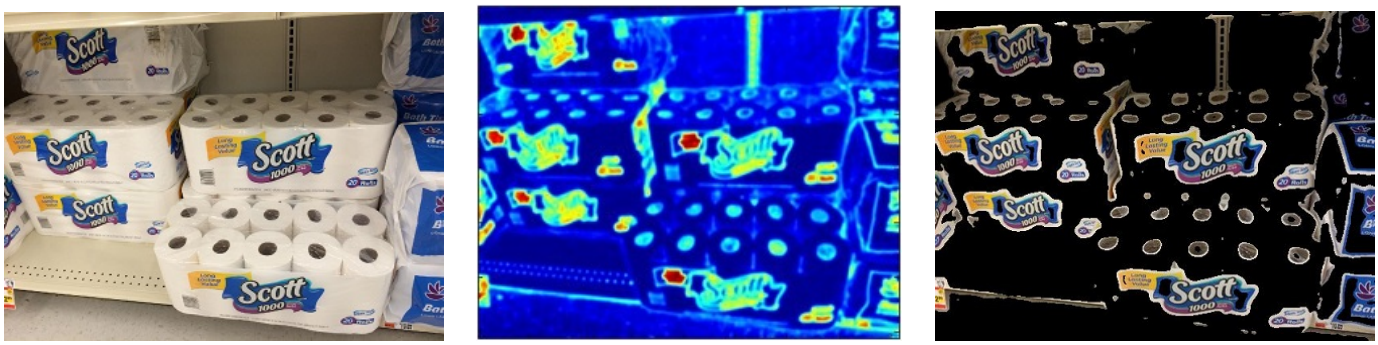
While the computational needs of the computer vision algorithms we employ are substantial, they are also heavily structured and ammenable to acceleration. The computation and memory management for sub-tasks, such as person-detection or recognizing sets of replicated objects on grocery shelves, can be heavily customized to yield both large performance and energy gains. We have developed FPGA-based solutions for these sub-tasks, although our designs could also translate to ASIC implementations.

In [8], the authors proposed a scalable solution for object detection using structured features - Histogram of Oriented Gradients. While these features are very lightweight in terms of hardware resources, the high miss-rate is cause for concern. On the other hand, a deep CNN would have better accuracy, but would struggle to meet stringent power and area constraints when being deployed onto a wearable platform. Our current work is focussed on reducing the detection threshold and using these detection outputs as region proposals to a shallow CNN. We use three convolutional (+ pooling) layers and three fully connected layers in our CNN. The training images were of size 64×128 with around 3000 positives and 6000 negatives. and we used stochastic gradient descent with a step-down approach to the learning rate. Figure 8(d) shows the output of our structured HOG custom hardware coupled with our shallow CNN.

B. Brain-like Architectures

The increasing prevalence of brain-inspired algorithms, such as saliency, and the dramatic rise in the use and utility of machine learning workloads has inspired research into architectures that more directly embody brain-like functionality. These new architectuures, such as IBM's TrueNorth [10] completely abandon the traditional, centralized von-Neumann architectures of general purpose computing for distributed, neuro-inspired computation models. While moving tasks to these new brain-like architectures generally requires significant rethinking and re-expression of the existing codebases, for tasks that were already modeling neural networks, such as many machine learning workloads, the impacts can be both rapid and profound.

IBM's TrueNorth chip is an architypical example of this new class of brain-like architectures. It consists of 4096 neurosynaptic cores arranged in a 2-D array occupying 4.3 cm^2 of area in a 28 nm low power CMOS process. A key advantage of this chip is that it consumes merely 65 mW of power while running a typical computer vision application [10]. Having accelerated some of the key computer vision models using custom fabrics like FPGAs and GPUs, we are now looking to map them onto TrueNorth. While our FPGA and GPU acceleration efforts have supported real-time performance levels and greatly increased power efficiency over traditional software



(a) Original image (b) Saliency Map (c) Thresholded Map

(a) Best RoI (b) SURF keypoints (c) Regions of Interest

Fig. 5. Saliency and SURF used to identify similar items.

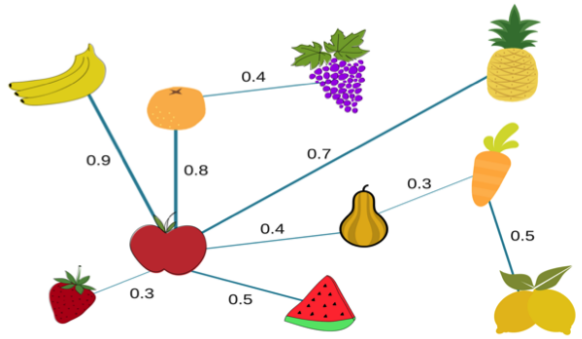


Fig. 6. Spatial relations exist between frequently co-occurring objects. These relationships can then be used as context cues to guide the recognition task.

approaches, moving to these new brain-like architectures offers the potential to push computation even closer to the sensing platform by easily operating within wearable power budgets.

For example, extracting a HOG-like feature vector for a given 64×128 Region of Interest (ROI) would require 23 cores consuming around 62 mW of power. To evaluate the fidelity of this feature, we trained an SVM using the INRIA dataset [11] and ran evaluations. Figure 9 (b) depicts the dot product output between the trained model and the TrueNorth HOG feature model, while Figure 9 (c) shows the thresholded detections when evaluated on a test image.

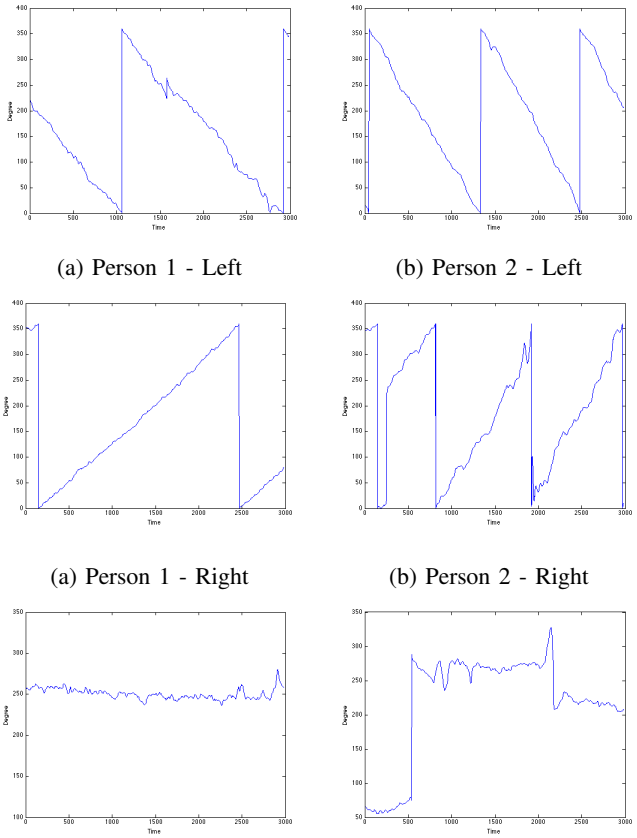
C. Emerging Devices

In addition to advances in novel architectures for vision, there are also new emerging technologies on the horizon that can potentially offer enable new visual computing paradigms. One such promising technology is weakly-coupled nano-oscillators. There is an existing body of work that shows how the analog functions of weakly-coupled oscillators can be used to provide distance-like metrics as a new analog primitive [12] and how sets of oscillators can implement analog convolutions and other useful high-level computational primitives. Recent advances in materials and device technologies offers the promise of nano-scale oscillators, such as hyperFET-based oscillators [13], that can implement these functions in extremely small area and power budgets.

Nano-oscillators are particularly intriguing for wearable vision applications. Early explorations indicate that arrays of hyperFET-based oscillators are small enough and consume sufficiently little power to be directly integrated into the image sensing chips. In addition to the benefits from offering higher-level analog primitives that improve computational efficiency relative over digital calculations, recent work has highlighted the large system-level benefits of moving early computation to the sensor chip [14], thereby limiting the losses incurred in moving data from the sensing to computing portions of wearable and mobile systems. Fully exploiting the potential of these nano-oscillators will, however, require substantial remapping efforts for existing computations. Our group has mapped a number of image pre-processing primitives onto coupled-oscillator arrays, and efforts are underway to map larger computations, such as HOG.

V. Conclusion

This work highlighted the efficacy of personal visual assist systems in our day to day activities. As technological advances



(a) Person 1 - Left (b) Person 2 - Left

(a) Person 1 - Right (b) Person 2 - Right

(a) Person 1 - Straight (b) Person 2 - Straight

Fig. 7. Temporal sensor information used for localization.

spur growth, more and more consumer products will become available.

VI. Acknowledgment

This work is supported in part by NSF Expeditions: Visual Cortex on Silicon CCF 1317560.

References

[1] Y. Bengio, "Learning Deep Architectures for AI." Now Publishers, 2009.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," pp. 436–444, May 2015.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *CoRR*, vol. abs/1503.03832, 2015. [Online]. Available: http://arxiv.org/abs/1503.03832

[4] N. Bruce, "An Information Theoretic Model of Saliency and Visual Search."

[5] I. Kokkinos and A. Yuille, "HOP: Hierarchical Object Parsing," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 802–809.

[6] S. Advani, B. Smith, Y. Tanabe, K. Irick, M. Cotter, J. Sampson, and V. Narayanan, "Visual co-occurrence network: Using context for large-scale object recognition in retail," in *Embedded Systems For Real-time Multimedia (ESTIMedia), 2015 13th IEEE Symposium on*, Oct 2015, pp. 1–10.

[7] B. E. Stein, T. R. Stanford, and B. A. Rowland, "The Neural Basis of Multisensory Integration in the Midbrain: its Organization and Maturation," *Hearing Research*, vol. 258, no. 1, pp. 4–15, 2009.

[8] S. Advani, Y. Tanabe, K. Irick, J. Sampson, and V. Narayanan, "A scalable architecture for multi-class visual object detection," in *Field Programmable Logic and Applications (FPL), 2015 25th International Conference on*, Sept 2015, pp. 1–8.

[9] Robust Multi-Person Tracking from Mobile Platforms. (Date last accessed 16-Sep-2016). [Online]. Available: https://data.vision.ee.ethz.ch/cvl/aess/dataset/

[10] F. Akopyan *et al.*, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 10, pp. 1537–1557, Oct 2015.

[11] INRIA person dataset. (Date last accessed 16-Sep-2016). [Online]. Available: http://pascal.inrialpes.fr/data/human/

[12] J. A. Carpenter, Y. Fang, C. N. Gnegy, D. M. Chiarulli, and S. P. Levitan, "An image processing pipeline using coupled oscillators," in *2014 14th International Workshop on Cellular Nanoscale Networks and their Applications (CNNA)*, July 2014, pp. 1–2.

[13] W. Y. Tsai *et al.*, "Enabling new computation paradigms with hyperfet - an emerging device," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 2, no. 1, pp. 30–48, Jan 2016.

[14] R. LiKamWa, Y. Hou, J. Gao, M. Polansky, and L. Zhong, "Redeye: Analog convnet image sensor architecture for continuous mobile vision," in *International Symposium on Computer Architecture*, 2016.

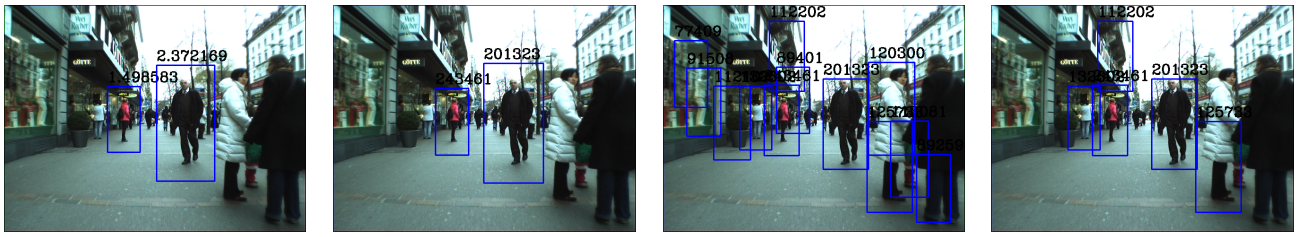
Peter A. Zientara received the B.S. degree in Computer Engineering from the York College of Pennsylvania in 2014 and is currently pursuing his Ph.D. in Computer Science and Engineering from the the Pennsylvania State University. His research interests are in smart sensors, hardware acceleration, and embedded vision systems.

Siddharth Advani received the B.S. degree in Electronics Engineering from the Pune University, India in 2005 and the M.S. degree in Electrical Engineering from the Pennsylvania State University in 2009. He is currently pursuing his Ph.D. in Computer Science and Engineering at the Pennsylvania State University. His research interests include embedded vision design for smart mobile applications targetted for domain-specific applications.

John (Jack) Sampson received the B.S. degree in Electrical Engineering and Computer Science from the University of California, Berkeley in 2002 and the Ph.D. in Computer Engineering from the University of California, San Diego in 2010. He is an Assistant Professor in the Department of Computer Science and Engineering at the Pennsylvania State University. His research interests include energy-efficient computing, architectural adaptations to exploit emerging technologies, and mitigating the impact of Dark Silicon.

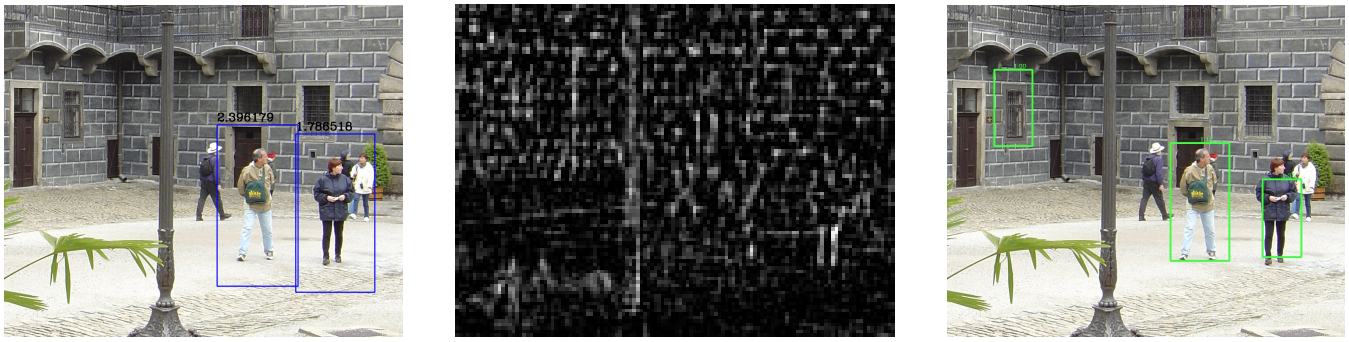
Vijaykrishnan Narayanan received the B.S. degree in Computer Science and Engineering from the University of Madras, India, in 1993 and the Ph.D. in Computer Science and Engineering from the University of South Florida, Tampa, USA, in 1998. He is a Professor of Computer Science & Engineering and Electrical Engineering at the Pennsylvania State University. His research interests include power-aware and reliable systems, embedded systems, nanoscale devices and interactions with system architectures, reconfigurable systems, computer architectures, network-on-chips, domain-specific computing.

Dr. Narayanan has received several awards including the Penn State Engineering Society Outstanding Research Award in 2006, IEEE CAS VLSI Transactions Best Paper Award in 2002, the Penn State CSE Faculty Teaching Award in 2002, the ACM SIGDA outstanding new faculty award in 2000, Upsilon Pi Epsilon award for academic excellence in 1997, the IEEE Computer Society Richard E. Merwin Award in 1996 and the University of Madras first rank in Computer Science and Engineering in 1993. He is currently the editor-in-chief of IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems. He has received several certificates of appreciation for outstanding service from ACM and IEEE Computer Society. He is a Fellow member of IEEE and ACM.



(a) OpenCV output (b) Hardware HOG output [8] (c) Reduced threshold (d) HOG-CNN output

Fig. 8. Coupling structured features with learned features. Image obtained from [9].



(a) OpenCV output (b) TrueNorth Dot Product (c) TrueNorth Detections

Fig. 9. Mapping HOG to True North. Image obtained from [11].