Performance Evaluation of Neuromorphic-Vision Object Recognition Algorithms

Rangachar Kasturi, Dmitry Goldgof, Rajmadhan Ekambaram
Department of Computer Science and Engineering
University of South Florida, Tampa, FL 33620, USA

Gill Pratt
Defense Advanced Research Projects Agency
675 North Randolph Street
Arlington, VA 22203-2114, USA

Eric Krotkov and Douglas D. Hackett
Griffin Technologies
PO Box 276
Wynnewood, PA 19096-0276, USA

Yang Ran and Qinfen Zheng
Image and Video Understanding Branch, Leidos
6301 Ivy Lane, Suite 200
Greenbelt, MD 20770

Rajeev Sharma
VideoMining Corporation
403 South Allen Street, State College, PA16801, USA

Mark Anderson, Mark Peot, Mario Aguilar
Information Sciences, Teledyne Scientific Company
5001 S. Miami Blvd.
Durham, NC 27703, USA

Deepak Khosla, Yang Chen, Kyungnam Kim
HRL Laboratories, LLC
3011 Malibu Canyon Rd,
Malibu, CA 90265, USA

Lior Elazary, Randolph C. Voorhies, Daniel F. Parks, Laurent Itti
University of Southern California
3641 Watt Way, HNB-07A
Los Angeles, CA 90089-2520, USA

***Abstract*— The U.S. Defense Advanced Research Projects Agency's (DARPA) Neovision2 program aims to develop artificial vision systems based on the design principles employed by mammalian vision systems. Three such algorithms are briefly described in this paper. These neuromorphic-vision systems' performance in detecting objects in video was measured using a set of annotated clips. This paper describes the results of these evaluations including the data domains, metrics, methodologies, performance over a range of operating points and a comparison with computer vision based baseline algorithms.**

Keywords—Neuromorphic vision, object detection, recognition, video analysis, performance evaluation

I. INTRODUCTION

DARPA's Neovision2 program's goal is to emulate the mammalian visual pathway by implementing advanced models and algorithmic emulations of the entire visual pathway - from retina to the visual cortex. The objective of the effort is to explore the potential for such neuromorphic systems to surpass regular engineering approaches by achieving significant improvements in size, weight, power, and performance. Five neuromorphic vision algorithms were developed under this program with an objective of detecting a specified set of objects typically seen in aerial images and video in urban environments. Manually annotated video clips were used to train these systems and a separate test set was used to quantify their performance. Two baseline algorithms based on the Deformable Part Model algorithm by Felzenszwalb [1] representing typical computer vision methods were also implemented to serve as a benchmark for comparing the performance of neuromorphic methods. In this short paper we provide a system-level description of three representative neuromorphic algorithms and a baseline algorithm, the datasets used for training and testing, performance measures, evaluation methodologies, and the results of these evaluations.

II. NEUROMORPHIC ALGORITHMS

We briefly describe three neuromorphic algorithms in this section.

A. HRL's Neuromorphic Image and Video Analysis (NIVA)

A team led by HRL Laboratories, LLC (HRL) included participants from New York University, Johns Hopkins University, MIT, Brown University, Yale University, Argon ST and Imagize, LLC. Our team successfully developed a real-time, accurate, and low-power neuromorphic solution for automated object recognition in images and videos [2,3]. Our system called NIVA (Neuromorphic Image and Video Analysis) is an autonomous object recognition system based on a visual cognition architecture (Fig. 1) that replicates the ventral (*what*) and dorsal (*where*) streams of the mammalian visual pathway by combining retinal processing (I), form- and motion-based object detection (II, III, IV) and convolutional neural nets based object classification (V) for unprecedented fast and accurate object recognition. This architecture is fundamentally different from raster-scan-like processing and engineered feature extraction of traditional computer vision and Automatic Target recognition (ATR) methods.

The front-end retinal-processing function (I) emulates the mammalian retina's unique properties of adaptability to illumination variation. In this program, our team developed a

hybrid EO/IR retinal camera (5.6Mpixel EO, 0.3Mpixel IR) that performs retina-like image enhancement of the incoming video in real time. The intermediate object detection module (II, III, IV) processes the output of the retinal stage to extract object regions. In its simplest form, this function computes tight bounding box regions around potential objects in the image via a fast, parallel, feed-forward step. It combines form-based detection that employs bio-inspired attention approaches to detect entities based on form (e.g., shape, color, intensity) and motion-based detection mediated by spatial attention. A typical example result of object detection is shown in Fig. 2. The final object classification module (V) then classifies the detected object regions into one of several pre-defined object class (e.g., car, person, and truck). The classification step uses a feed-forward multi-layered convolutional neural net [4] approach that is fast, efficient and inherently parallel. These networks are inspired by the architecture of mammalian visual pathways, and consist of alternating stages of layers of simple cells followed by complex cells. As the visual information goes through the stages of processing, the extracted features become more global and invariant. Our team mapped the above neuromorphic architecture and algorithms to commercially available off-the-shelf hardware. The dynamic power requirement for the complete NIVA system that includes retinal camera, object detection and classification was measured by DARPA at 21.7 W. This corresponds to an equivalent energy consumption of 5.4 nJ per bit.

In summary, HRL's NIVA is a real-time, accurate, and ultra-low power autonomous video object recognition system that is suitable for on-board and off-board processing of live or recorded images and videos. It is agnostic to both data resolution and type and can be hosted on off-the-shelf hardware, making it practical and applicable for mainstream use.

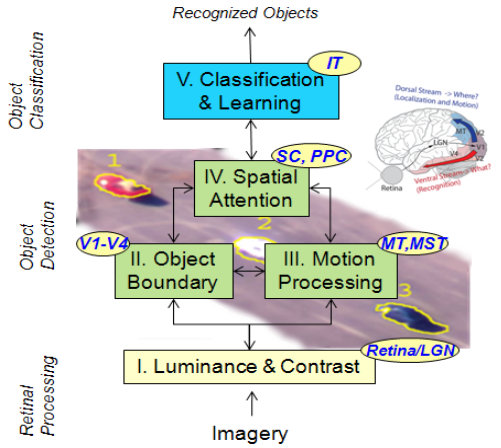

Fig. 1. HRL's NIVA integrated visual cognition architecture that emulates mammalian dorsal and ventral visual functions. Five functional components I-V unify the state of the art in retinal processing, object detection and classification. (LGN: Lateral Geniculate Nucleus, V1-V4: Visual Cortex, MT: Middle Temporal, MST: Medial Superior Temporal, SC: Superior Colliculus, PPC: Posterior Parietal, IT: Infero-Temporal).

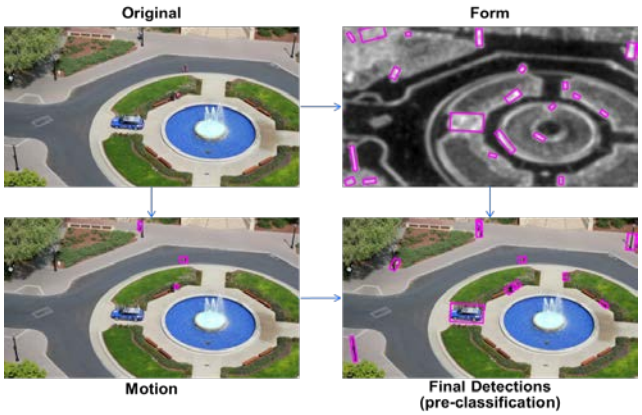

Fig. 2. HRL's NIVA object detection (II, III, IV) example by fusion of form and motion pathways. Input image (top left) is processed by separate and parallel pathways for form (top right) and motion (bottom left). The detections for each pathway are fused (bottom right) and then subsequently classified.

B. Neuromorphic Enhancement and Hierarchical Decomposition.

Teledyne Scientific, with a team of academics (Duke, Georgia Tech,, Harvard, Penn State, Vanderbilt, University of Ljubjana and York University) and industry partners (Sarnoff and Markury Scientific), developed and demonstrated an object recognition solution that achieved real-time performance on complex object recognition tasks. Our approach to achieving this performance was the development of an architecture that emulates the neuromorphic and computational principles present in the mammalian visual pathway to support tactically relevant accuracy and unsurpassed robustness (see Fig. 3).

Our model for early stages of processing in the retina and lateral geniculate nucleus (LGN) is based on the pioneering work of Stephen Grossberg [5,6]. The early vision system model conditions the image by discounting illumination effects, enhancing color contrast and mitigating imaging artifacts [7,8,9] such as atmospheric transmission/contrast loss. We model processing in the primary visual cortex (visual area V1) as a multi-scale, multi-orientation log Gabor filtering process that delineates object boundaries that serve as the foundation for shape-based recognition. Next, we utilize mechanisms of multi-scale spatial facilitation, inspired by processes in V2, to enhance and complete edges, while minimizing the impact of noise. We employ a model of saliency proposed by [10] to allocate resources and processing in V2 and beyond ("Priming" box in Fig. 3). This attention model selects image regions that are the most likely to contain objects of interest based on a self-information metric and image statistics. Processing akin to that found in area V4 of the visual system is captured by the learned hierarchy of parts (LHoP) algorithm [11]. LHoP hierarchically decomposes objects into libraries of reusable parts. At each level of the hierarchy, objects are described using combinations of parts from a previous level. Activations across the hierarchy of parts is used to define a rich feature set for attended objects which is analyzed at the IT stage by an ARTMAP-based pattern recognition stage. ARTMAP [12] is modeled after computational principles in cortical areas involved in learning and retrieval of patterns for stored memories and associations. In our system, ARTMAP serves as the final stage to assign

object categories (labels) to the activated features in the hierarchy of parts stage (V4).

We adopted a novel approach for training our system. Rather than using hand-labeled imagery, we learned part libraries from 2D projections of 3D models of figures and vehicles drawn in Google Sketchup.

Fig. 4 shows intermediate results from our algorithm. First, our retinal and LGN algorithms work together to enhance scene contrast, level dynamic range and encode color contrast information across three color channels (left). Next, we extract edges at multiple scales in each image and aggregate the results (middle left). For each patch identified by our attention algorithm, we use a hierarchical library of flexible parts (middle right) to encode each object (right). Once the object is encoded using the part library, we use the vector of part activations with the ARTMAP classifier to classify the object.

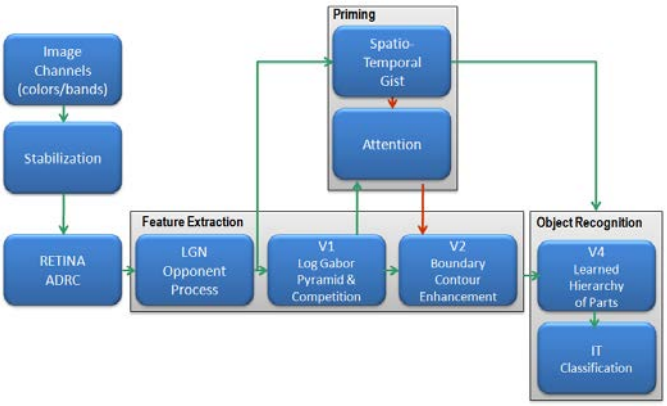

Fig. 3. Teledyne's Neuromorphic Enhancement and Hierarchical Decomposition object recognition architecture.

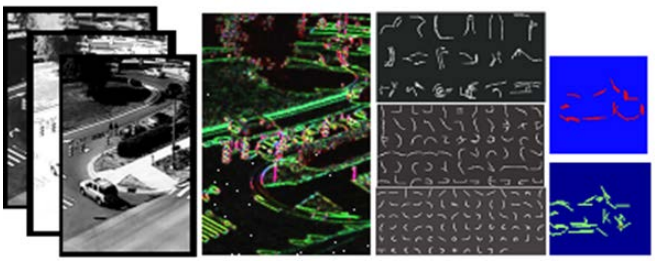

Fig. 4. Processing pipeline for Teledyne's neuromorphic enhancement and hierarchical decomposition.

C. USC's Cognitive Scene Understanding System.

The team led by The University of Southern California (USC) included participants from MIT, Brown University, Caltech, Penn State University, Duke University, UC Berkeley, Arizona State University, Queen's University and Imagize. The developed algorithm implements many of the known components of human visual intelligence [13], including rapid processing of the gist of a scene, visual attention, object recognition, multiple-object tracking, and a spatial working memory (Fig. 5).

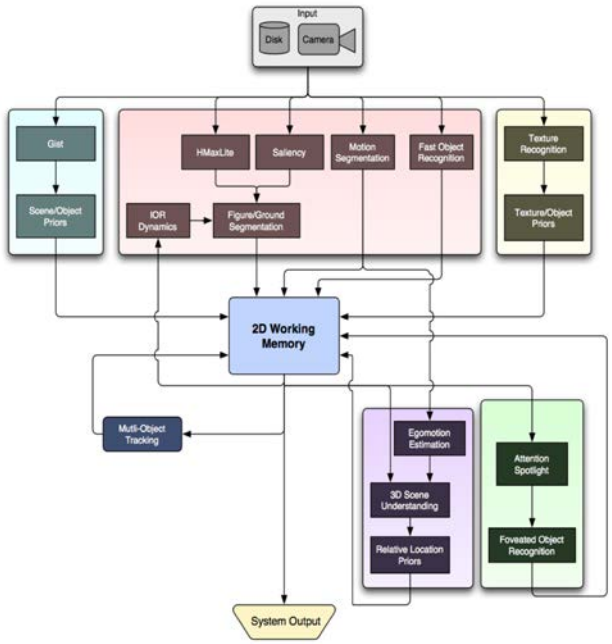

Fig. 5. Flow diagram of the USC algorithm.

Briefly, the algorithm operates as follows: Input video is first subjected to a coarse but rapid feature extraction stage that operates over the entire visual field in parallel. Features extracted include gist features (which provide a prior as to where and which objects of interest might appear based on global statistical analysis of the image) [14], visual saliency features [15], motion features, early-stage HMAX features [16], and fast but crude object recognition features [17]. These features all combine into guiding an attention spotlight towards the image locations more likely to contain objects of interest. As the spotlight shifts over these locations in turn, each is coarsely segmented and sent to the central working memory. This memory maintains a list of known (and tracked) entities (Fig. 6). It implements a probabilistic Bayes filter, such that, on every new video frame, it guesses where each tracked entity might have moved, and then confirms or corrects these guesses using the new incoming data. The filter also builds over time a belief distribution over object categories for each tracked entity. A parallelized object recognition pipeline based on the HMAX algorithm [16] attempts to recognize each of the known entities, augmenting their representation with a probabilistic class label as soon as sufficient confidence has built up about this label. Finally, confidently attended to, recognized, and tracked objects in the 2D images are further validated against an incrementally built 3D world model in working memory.

The component algorithms can be composed in different arrangements thanks the development of the Neuromorphic Robotics Toolkit, a Blackboard-based message passing framework that operates on clusters of CPU, GPU, and FPGA hardware. Source code is freely available from the authors.

Fig. 6. Examples of detections with the USC algorithm on Heli clips. Black boxes have been attended to but classified as background and are eliminated from final output. Blue: cars; Yellow: humans; Pink: bicycles.

III. Baseline Computer Vision Algorithm

The neuromorphic vision algorithms' performance in recognizing several object classes in video were compared against a baseline computer vision object recognition algorithm. The baseline algorithm is based on deformable part models [1]. The algorithm is implemented on a CUDA (Compute Unified Device Architecture) platform for GPUs trained to recognize objects. Each object class for each data domain (described in the next section) was trained using hundreds of positive and negative examples. By focusing on foreground objects after background subtraction, substantial speedup and power savings were achieved. Results are presented in Section V.

IV. Data Domains and Annotation Methodology

To evaluate the performance of these algorithms, video from three data domains were captured. Two datasets of 100 video clips each were captured from two data domains which are labeled as Tower and Helicopter. The tower dataset is filmed from a fixed camera on top of the Stanford University Hoover tower and the Helicopter dataset is filmed from a helicopter flying over Los Angeles. In both cases, the 1080p video imagery is converted to 8 bit PNG frames for analysis. A third dataset from DARPA TAILWIND (Tactical Aircraft to Increase Long Wave Infrared Nighttime Detection) program is captured from an airplane operating at two different altitudes. Characteristics of these datasets are summarized in Fig. 7. Human annotators at VideoMining Corporation labeled ten object classes (Boat, Car, Container, Cyclist, Helicopter, Person, Plane, Tractor-Trailer, Bus, and Truck) in these datasets. Each object is enclosed in an oriented rectangle that best encompasses the object. Ten percent of the datasets are annotated by two independent annotators and their outputs are compared to assess quality and consistency of annotations; significant differences between the two sets trigger a review of the annotation process to assure annotation quality. Each annotated dataset is divided into training and test sets. The Tower and Helicopter training sets are available for download from http://ilab.usc.edu/neo2/dataset/. The test datasets are also available upon request. Additional details of the imagery, the platforms used to collect them, and the annotation guidelines are also available at the same site.

	Tower	Helicopter	TAILWIND
Resolution	1920 x 1088	1920 x 1080	3660 x 2748
Size (megapixel)	2	2	10
Size (megabits)	50	50	241
Ground sample distance (pixel/m)	30 - 40	25 – 40	4 – 8 High alt. 20 – 40 Low alt.
Frame rate (fps)	30	30	5
Bandwidth (Gb/s)	1.5	1.5	1.0

Fig. 7. Representative images from the three data domains and their characteristics.

V. Performance Evaluation

Since the data source characteristics and perhaps the corresponding Neovision2 systems are different for the three data domains, the performance of the systems are evaluated and reported separately for each domain. The evaluation methodology is based on the VACE (Video Analysis for Content Extraction) performance evaluation described in [18]. The performance of the system in detecting each of the object classes is evaluated independent of other classes using Normalized Multiple Object Thresholded Detection Accuracy (NMOTDA, defined later in this section), missed detects and false positives.

A Neovision2 object is denoted by a class ID and a bounding box (4 corners in (x, y) pixels in a frame, where (0,0) is the top left corner of the frame.). The following notations are used:

- $G_i^{(t)}$ denotes the ith ground-truth object in frame t.
- $D_i^{(t)}$ denotes the ith detected object in frame t.
- $N_G^{(t)}$ and $N_D^{(t)}$ denote the number of ground-truth objects and the number of detected objects in frame t, respectively.
- N_{frames} is the number of frames in the video sequence.

A. Spatial Overlap Threshold for Scoring Object Detection

We use the spatial overlap between a pair of ground-truth object and a system output object for detection scoring. For a pair of $G_i^{(t)}$ and $D_i^{(t)}$, the overlap ratio is calculated as:

$$\text{Overlap\_Ratio} = \frac{\left| G_i^{(t)} \cap D_i^{(t)} \right|}{\left| G_i^{(t)} \cup D_i^{(t)} \right|} \quad (1)$$

An oriented ground truth box is converted to a corresponding vertical box (minimal vertical rectangular envelope) before it is

compared with the system output using Eq. 1. An object instance is considered as detected if the *Overlap_Ratio* is greater than or equal to a preset value of 0.2.

In each frame there can be multiple instances of the same object class, for example, N cars represented by N ground truth boxes. A Neovision2 car detection system may output M boxes for this image. Ideally M is equal to N but in general it may be less or more than N. We must first assign the system output boxes to the corresponding ground truth boxes in a one-to-one fashion. Even when N=M there are N! possible combinations of such assignments. When N is not equal to M we have unmatched boxes. Our assignment algorithm finds the matching pairs that maximize the detection performance measure using a numerical search algorithm that guarantees arriving at one optimal solution [19]. Once these unique assignments are made we apply the *Overlap_Ratio* threshold for each matched pair. If the ratio is greater than the threshold the object is considered detected and a score of 1 is assigned. Each ground truth box that has no matching pair or a match with a spatial overlap of less than 20% is a Missed Detect. Similarly, each system output box which has no matching pair or a match with a spatial overlap of less than 20% is a False Positive.

B. Performance Measure for Sequence

Performance measure for a sequence is obtained by matching the ground truth and system output at each frame from that sequence. Aggregating these overall images in the entire sequence, we get the Normalized Multiple Object Thresholded Detection Accuracy, NMOTDA given by

$$NMOTDA = 1 - \frac{\sum_{t=1}^{N_{frames}} (c_m m^{(t)} + c_f f^{(t)})}{\sum_{t=1}^{N_{frames}} N_G^{(t)}} \qquad (2)$$

where c_m and c_f are the cost functions for missed detects and false positives, $m^{(t)}$ and $f^{(t)}$ are the number of missed detects and false positives in frame t in the sequence. The summations are carried out over all evaluated frames. In Neovision2 evaluations, the cost functions c_m and c_f are set equal to 1. This is a sequence-based measure which penalizes false detections, missed detections and object fragmentation. Note that maximizing NMOTDA for the sequence is the same as finding the optimal assignment of ground truth boxes to system output boxes at each frame image. NMOTDA has a range of 1 (best) to -∞.

C. Performance Measure for Domain

In each data domain, multiple sequences are used for evaluation. The summary of the performance over the entire dataset, i.e., over all the video clips, is calculated using *Weighted* NMOTDA (WNMOTDA). WNMOTDA is calculated for each class separately and is given by,

$$WNMOTDA_i = \frac{\sum_{j=1}^{NVC} NMOTDA_{ij} * NGT_{ij}}{\sum_{j=1}^{NVC} NGT_{ij}} \qquad (3)$$

where $WNMOTDA_i$ is the NMOTDA for class i (boat, car etc.) calculated over all the video clips, $NMOTDA_{ij}$ is the NMOTDA measure calculated for class i in video clip j, NVC is the number of video clips, and NGT_{ij} is the number of ground truth objects of class i in video clip j.

Finally an Average WNMOTDA score is generated for all object classes for each domain using Eq. (3) by ignoring the class labels. However, before scoring identical boxes are merged into one. Overlapped boxes (if *Overlap_Ratio* is over 20%) are merged into one and their union is used instead.

VI. Results of Performance Evaluation

The results of formal performance evaluation using the test set are shown in Figs. 8 and 9. The three neuromorphic systems are labeled as Team A (HRL), Team B (USC), and Team C (Teledyne) in these figures. These results are summarized as WNMOTDA scores for each domain in Fig. 8. Each symbol represents the WNMOTDA value of an object class, e.g., Truck. These results show that neuromorphic vision systems' performance is on par with the baseline computer vision algorithm. In addition to these software performance evaluations, neuromorphic systems' energy consumption was measured during live processing. These results are summarized in Fig. 9 using Average WNMOTDA scores. These results show that neuromorphic systems' energy consumption is four orders of magnitude lower than conventional systems.

This research has demonstrated that neuromorphic computing has the potential to enable autonomous and low-power capability on board a UAV such that only relevant results of interest need to be transmitted to the ground station, thereby reducing both the requirements for data bandwidth and analyst man power.

VII. Summary and Conclusions

In this paper, we presented three neuromorphic object recognition algorithms built to emulate the mammalian visual pathways. The performance of each algorithm was evaluated using annotated test sets from three domains. The results were compared with the baseline computer vision algorithm. These demonstrate that neuromorphic systems perform on par with baseline computer vision algorithms in object recognition tasks. Energy consumption tests confirmed their substantial advantage over traditional approaches.

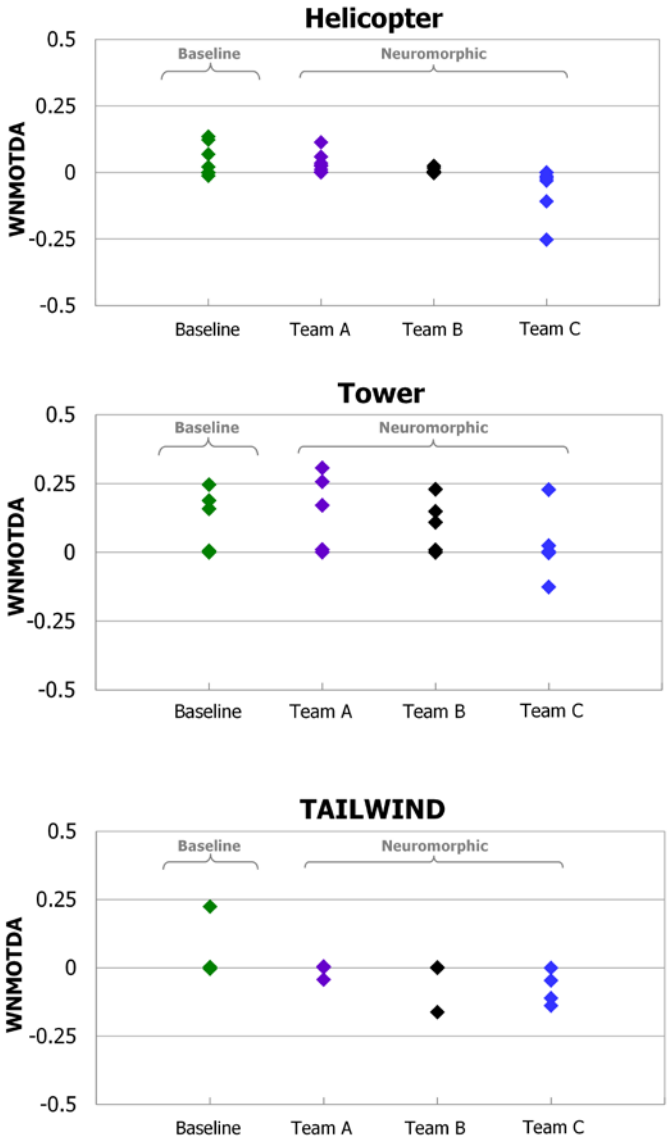

Fig. 8. Results of performance evaluation of neuromorphic and computer vision baseline algorithms on Helicopter, Tower and TAILWIND datasets. Each symbol represents the WNMOTDA value of an object class, e.g., Truck.

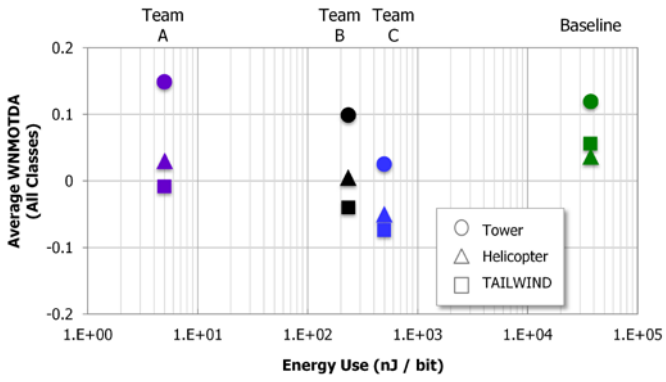

Fig. 9. Energy consumption of neuromorphic and baseline computer vision algorithms.

REFERENCES

[1] P. Felzenszwalb, R. Girshick, and D. McAllester, "Cascade Object Detection with Deformable Part Models," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[2] D. Khosla, Y. Chen, K. Kim, S. Cheng, A. Honda, L. Zhang, "A neuromorphic system for object detection and classification", *Proc. SPIE 8745, Signal Processing, Sensor Fusion, and Target Recognition XXII*, 87450X, 2013.

[3] D. Khosla, Y. Chen, D. Huber, et al., "Real-time low-power neuromorphic hardware for autonomous object recognition", *Proc. SPIE 8713, Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications X*, 871313, 2013.

[4] Y. LeCun, K. Kavukcuoglu, and C. Farabet "Convolutional Networks and Applications in Vision", *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, 253-256, 2010.

[5] Ellias, S.A. and Grossberg, S. (1979) "Pattern formation, contrast control, and oscillations in the short memory of shunting on-center off-surround networks", Biological Cybernetics, 20, pp. 69-98.

[6] Grossberg, S. (1988) "Neural Networks and Natural Intelligence", Cambridge, MA: MIT Press

[7] Aguilar, M., J. Grater, R. Tarquini, J. Johnson. (2006) "Enhancement, Fusion, and Visualization of Third Generation FPA Imagery." In Proc. of the Meeting of the Military Sensing Symposia (MSS), Specialty Group on Passive Sensors. Orlando, FL, 13-17 February 2006.

[8] Aguilar, M., Fay, D.A., Ireland, D.B., D.A., Racamato, J.P., Ross, W.D., and Waxman, A.M (1999). "Fusion of Low-Light Visible and LWIR for Color Night Vision." In Proc. of the 2nd Int. Conf. on Information Fusion, Sunnyvale, CA.

[9] Aguilar, M., Fay, D.A., Ross, W.D., Waxman, A.M., Ireland, D.B. and Racamato, J.P. (1998). "Real-Time Fusion of Low-Light CCD and Uncooled IR Imagery for Color Night Vision." In Proc. of SPIE Enhanced and Synthetic Vision, vol. 3364, Orlando, FL.

[10] Bruce, N.D.B., Tsotsos, J.K., Saliency, Attention, and Visual Search: An Information Theoretic Approach, Journal of Vision 9:3, p1-24, 2009, http://journalofvision.org/9/3/5/, doi:10.1167/9.3.5.

[11] Fidler, S. and Leonardis, A. (2007). "Towards scalable representations of object categories: Learning a hierarchy of parts", Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'07, pp. 1-8.

[12] Carpenter, G., Grossberg, S. and Reynolds. J.H. (1991), "ARTMAP: Supervised real-time learning and classification of nonstationary data by self-organizaing neural network.", *Neural Networks*, 4(5), pp. 565-588.

[13] V. Navalpakkam, L. Itti, Modeling the influence of task on attention, Vision Research, Vol. 45, No. 2, pp. 205-231, Jan 2005.

[14] C. Siagian, L. Itti, Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 2, pp. 300-312, Feb 2007.

[15] L. Itti, C. Koch, E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 11, pp. 1254-1259, Nov 1998.

[16] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio (2007). Robust object recognition with cortex-like mechanisms. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29(3), 411-426.

[17] L. Elazary, L. Itti, A Bayesian model for efficient visual search and recognition, Vision Research, Vol. 50, No. 14, pp. 1338-1352, Jun 2010.

[18] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 2009.

[19] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense & sparse linear assignment problem" *Computing*, 38, 325-340, 1987.