

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

CROWDSTRIKE, INC.,

Petitioner

v.

SKYSONG INNOVATIONS, LLC,

Patent Owner

Inter Partes Review Case No. IPR2025-01399

U.S. Patent No. 11,775,831

DECLARATION OF Dr. ELI SABER, Ph.D.

TABLE OF CONTENTS

I.	EDUCATIONAL BACKGROUND AND QUALIFICATIONS	6
II.	METHODOLOGY AND MATERIALS CONSIDERED	11
III.	LEGAL FRAMEWORK.....	14
IV.	OPINION	21
A.	THE '831 PATENT	21
B.	SUMMARY OF THE '831 PATENT'S PROSECUTION HISTORY	24
C.	CLAIM CONSTRUCTION	25
D.	LEVEL OF A PERSON OF ORDINARY SKILL IN THE ART	25
V.	BACKGROUND OF THE TECHNOLOGY	27
A.	CONVOLUTIONAL NEURAL NETWORKS.....	27
B.	MAX POOLING IN CNNs	29
C.	RECTIFIED LINEAR UNIT IN CNNs	30
D.	MOST SIGNIFICANT BITS.....	31
E.	IMPROVING EFFICIENCY IN DEEP LEARNING	32
F.	ADAPTIVE PRECISION.....	33
VI.	OPINION: GROUND 1	34
A.	OBVIOUSNESS OVER UJIE AND MOONS	34
1.	<i>Overview of Ujiie.....</i>	<i>34</i>
2.	<i>Overview of Moons.....</i>	<i>38</i>
3.	<i>Claim 1(Pre): "One or more non-transitory computer-readable storage media storing instructions which, when executed by at least one processor, cause the at least one processor to perform operations comprising:"</i>	<i>39</i>
4.	<i>Claim 1(a): "in one or more layers of a convolutional neural network (CNN), performing a first iteration that includes computing a value based on a first set of most significant bits (MSBs) for each of a plurality of data sets;"</i>	<i>41</i>
a.	Motivation to Combine Ujiie and Moons	51
5.	<i>Claim 1(b): "examining a first set of values computed for the plurality of data sets in the first iteration to determine</i>	

	<i>whether a maximum value is present among the first set of values;”</i>	64
6.	<i>Claim 1(c): “responsive to identifying the maximum value, performing a full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value; and”</i>	66
7.	<i>Claim 1(d): “propagating the full precision computation of the value to a subsequent layer of the CNN.”</i>	68
8.	<i>Claim 10: The one or more non-transitory computer-readable storage media of claim 1, wherein the CNN is employed to analyze an image.</i>	72
9.	<i>Claim 11: The one or more non-transitory computer-readable storage media of claim 1, wherein:</i>	72
a.	<i>Claim 11(a): “the first iteration computes a value that approximates the full precision computation of the value; and”</i>	73
b.	<i>Claim 11(b): “the full precision computation is performed on the data set that includes less data than the plurality of data sets.”</i>	73

VII. OPINION: GROUNDS 2 AND 3 75

A.	OBVIOUSNESS OVER UJIE, MOONS, AND KAUL, OR ALTERNATIVELY OBVIOUS OVER UJIE, MOON, KAUL AND ISHII	75
1.	<i>Overview of Kaul</i>	75
2.	<i>Overview of Ishii</i>	75
3.	<i>Claim 2: The one or more non-transitory computer-readable storage media of claim 1, further comprising: responsive to determining that the first set of values are the same, performing, by the at least one processor, a second iteration that includes computing the value based on a second set of MSBs for each of the plurality of data sets, the second set of MSBs being larger than the first set of MSBs.</i>	77
a.	<i>Ujie-Moons in View of Kaul</i>	77
i.	<i>Motivation to Combine Ujie-Moons and Kaul</i>	82
b.	<i>Ujie-Moons-Kaul in View of Ishii</i>	88
i.	<i>Motivation to Combine Ujie-Moons-Kaul and Ishii</i>	92
4.	<i>Claim 3</i>	95

a.	Claim 3(a): “The one or more non-transitory computer-readable storage media of claim 2, further storing instructions, which, when executed by the at least one processor, cause the at least one processor to perform operations comprising: examining a second set of values computed for the plurality of data sets in the second iteration to determine whether the maximum value is present among the second set of values; and”	95
b.	Claim 3(b): “responsive to identifying the maximum value among the second set of values, performing, by the at least one processor, the full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value in the second iteration.”	97
5.	<i>Claim 4: The one or more non-transitory computer-readable storage media of claim 2, wherein the computing in each of the first iteration and the second iteration employs a convolution and a pooling.</i>	<i>99</i>
6.	<i>Claim 5: The one or more non-transitory computer-readable storage media of claim 4, wherein the convolution is a NxN convolution, where N is any integer.</i>	<i>100</i>
7.	<i>Claim 6: The one or more non-transitory computer-readable storage media of claim 4, wherein the pooling is a NxN pooling, where N is any integer.</i>	<i>101</i>
8.	<i>Claim 8: The one or more non-transitory computer-readable storage media of claim 2, wherein at least one of the first iteration and the second iteration is performed with a precision less than that of the full precision computation.</i>	<i>102</i>
9.	<i>Claim 9: The one or more non-transitory computer-readable storage media of claim 8, wherein the precision is 8-bit precision.</i>	<i>103</i>

VIII. OPINION: GROUNDS 4 AND 5 108

A.	OBVIOUSNESS OVER UJIE, MOONS, KAUL, AND RAVINDRAN, OR ALTERNATIVELY, OBVIOUS OVER UJIE, MOONS, KAUL, ISHII, AND RAVINDRAN	108
1.	<i>Overview of Ravindran.....</i>	<i>108</i>

2.	<i>Claim 7: The one or more non-transitory computer-readable storage media of claim 4, wherein the convolution is a 3x3 convolution, and the pooling is a 2x2 pooling.</i>	109
a.	Motivation to Combine Ujiie-Moons-Kaul-Ishii and Ravindran	111
IX.	CONCLUSION	116

I, Eli Saber, declare as follows:

1. I am over the age of 21 and am competent to make this declaration. I make this declaration based on facts and matters within my own knowledge and on information provided to me by others, and, if called as a witness, I could and would competently testify to the matters set forth herein.

I. EDUCATIONAL BACKGROUND AND QUALIFICATIONS

2. My qualifications are set forth in my curriculum vitae, which is attached as Appendix A. The following is a brief summary of my relevant qualifications and professional experience.

3. I am a professor in the Electrical and Microelectronic Engineering Department and the Chester F. Carlson Center for Imaging Science at the Rochester Institute of Technology (RIT) in Rochester, New York. I also serve as the Director of the Image, Video, and Computer Vision Laboratory. I joined RIT's full-time faculty in 2004. I have over thirty-five (35) years of industry and academic experience, thirty (30) of which are in the fields of Computer Vision, Digital Image/Video Processing, Pattern Recognition and Machine Learning.

4. I received a Ph.D. in Electrical Engineering from the University of Rochester in 1996. My concentration was on Signal/Image/Video Processing, Pattern Recognition, and Computer Vision. I received a Master's Degree in Electrical Engineering from the University of Rochester in 1992, and I received a

Bachelor's Degree in Electrical and Computer Engineering, summa cum laude, from the State University of New York at Buffalo in 1988.

5. Before becoming a full-time professor, I worked for Xerox Corporation for 16 years, from 1988 to 2004. During my years at Xerox, I was responsible for delivering color management, image processing innovations, architectures, and algorithms; xerographic sub-systems for a variety of color products; and control systems for toner production facilities. One of several roles I held at Xerox was Advanced Development Scientist and Manager. In that capacity, I established the Advanced Design Laboratory—an Imaging/Xerographics lab—and provided technical and managerial leadership for the Electrical, Imaging and Xerographics Department. In another role, as Product Development Scientist and Manager, I led the research and development of image quality metrics for various product platforms. I also led the Image Science, Analysis, and Evaluation area, with 12 to 15 direct reports and a budget of approximately \$2 million.

6. From 1997 until 2004, I was an adjunct faculty member in the Electrical Engineering Department of RIT and in the Electrical and Computer Engineering Department of the University of Rochester. I was responsible for teaching undergraduate and graduate coursework in signal, image, and video processing; pattern recognition; and communications. Additionally, I performed research in

multimedia applications, pattern recognition, image understanding, and color engineering.

7. Since joining RIT full-time in 2004, I have been responsible for teaching a variety of undergraduate and graduate courses in digital signal processing, digital image processing, digital video processing, engineering analysis, random signal and noise, advanced engineering mathematics, matrix methods, pattern recognition, communications, modern control theory, and linear systems.

8. I proposed, developed and introduced the Digital Video Processing course and the “Advanced Neural Network and Deep Learning: Part 1” courses; and reestablished the Pattern Recognition course at the Electrical Engineering Department at RIT. These graduate courses were instrumental in providing MS/PhD students with the proper foundation and knowledge in all aspects of the subject areas to perform in-depth research in the field and continue to advance the state of the art.

9. The Digital Video Processing course is designed to provide a thorough and comprehensive coverage of the field. Topics covered include the mathematical representation of digital video including motion and time-varying image formation models, spatio-temporal sampling/reconstruction, sampling structure conversions, motion estimation using optical flow, block-based, pel-recursive, Bayesian and point correspondence type methodologies, segmentation, motion tracking, filtering, restoration, super-resolution and compression.

10. The Pattern Recognition course, on the other hand, covers pertinent concepts in Bayes decision theory, linear and nonlinear classifiers, artificial neural networks, feature selection/generation/extraction and various types of clustering algorithms and techniques.

11. The “Advanced Neural Networks and Deep Learning: Theory, Mathematics and Algorithms” is intended to provide an in-depth comprehensive study of the theory, mathematics and algorithms for advanced neural networks and their usage in deep learning. It is primarily targeted for upper graduate students that are planning to pursue advanced studies/research in the field. The course begins with an in-depth study of the mathematical methodologies for neural networks and optimization (zero-, first- and second-order) techniques. It then progresses to provide a thorough study of shallow and deep neural networks architectures and algorithms, activation functions, regression, two- and multi-class classification, unsupervised learning, feature engineering/selection, backpropagation along with applications to a variety of engineering problems.

12. I am currently working to develop a follow-on course (Part 2) to the Advanced Neural Networks and Deep Learning course described above. This course is intended to cover convolutional and recurrent style neural networks, principles of feature engineering, kernel methods, tree-based learners, attention

mechanisms/transformers, reinforcement learning and generative adversarial networks. This course is targeted to debut in Fall 2026.

13. I am also currently working to revamp the graduate course titled: “Image and Video Compression”. The updated version of this course will start by providing a brief history of video coding and video coding standards and then progress to discussing video basics, formats and fundamentals of video compression. It will then progress to cover JPEG, JPEG 2000, H.261, MPEG-1, MPEG-2, H.263, MPEG-4 visual, H.264 and H.265 and if time permits, AV-1 and H.266. The course is intended to debut in Fall 2027.

14. My current research focuses on image/video processing for multimedia, military & biomedical applications; computer vision and three-dimensional scene reconstruction; and machine learning for object tracking/remote sensing applications. As a principal investigator or co-principal investigator, I have acquired research funding in excess of \$5 million since joining RIT and have managed multiple government grants from the Department of Defense as well as several yearly corporate grants from Hewlett-Packard. In 2012, I was awarded the Prestigious Trustees Scholarship, the highest award at RIT with regard to research recognition.

15. I am the author or co-author of 40 peer-reviewed journal publications. I have also authored or co-authored 103 conference and workshop publications and

a book entitled Advanced Linear Algebra for Engineers with MATLAB, published by CRC Press in February 2009, and am a named inventor on multiple U.S. and foreign patents.

16. I am a senior member of the Institute of Electrical and Electronic Engineers and a member of the IEEE Signal Processing Society, the Electrical Engineering Honor Society, Eta Kappa Nu.

17. In my consulting career, I have provided both deposition testimony and trial testimony.

18. I am being compensated in this matter at my customary rate of \$600/hour. My compensation in this matter is not based on the substance of my opinions or on the outcome of this matter. I have no financial interest in the Petitioner, or in the Patent Owner. I have no financial interest in the outcome of this matter or in any litigation involving the '831 Patent.

II. METHODOLOGY AND MATERIALS CONSIDERED

19. I have relied upon my education, knowledge and experience with the technology of neural networks, as well as the other materials as discussed in this declaration in forming my opinions.

20. For this work, I have been asked to review U.S. Patent No. 11,775, 831 ("'831 Patent") (EX1001) including the specification and claims, and the '831 Patent's prosecution history ("'831 File History") (EX1002). In developing my

opinions relating to the '831 Patent, I have considered the materials cited or discussed herein, including those listed below.

Exhibit	Description
1001	U.S. Patent No. 11,775, 831 (“’831 Patent”)
1002	File History of the '831 Patent (“’831 File History”)
1004	Takayuki Ujiie <i>et al.</i> , <i>Approximated Prediction Strategy for Reducing Power Consumption of Convolutional Neural Network Processor</i> , 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (2016) (“Ujiie”)
1005	Bert Moons <i>et al.</i> , <i>Energy-Efficient ConvNets Through Approximate Computing</i> , IEEE Winter Conference on Applications of Computer Vision (2016) (“Moons”)
1006	Himanshu Kaul <i>et al.</i> , <i>A 21.5M-Query-Vectors/s 3.37nJ/Vector Reconfigurable k-Nearest-Neighbor Accelerator with Adaptive Precision in 14nm Tri-Gate CMOS</i> , 2016 IEEE International Solid-State Circuits Conference (2016) (“Kaul”)
1007	U.S. Patent Application Publication No. 2016/0259995 A1 to Ishii <i>et al.</i> (“Ishii”)
1008	U.S. Patent Application Publication No. 2016/0259994 A1 to Ravindran <i>et al.</i> (“Ravindran”)
1009	Declaration of June Munford (“Munford Dec.”)
1010	Declaration of Gordon MacPherson (“MacPherson Dec.”)
1011	Matthieu Courbariaux <i>et al.</i> , <i>Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to +1 or -1</i> , arXiv: 1602.02830v3 [cs.LG] (March 17, 2016) (“Courbariaux”)
1013	Yann LeCun <i>et al.</i> , <i>Gradient-Based Learning Applied to Document Recognition</i> , Proceedings of the IEEE (1998) (“LeCun I”)
1014	Jiuxiang Gu <i>et al.</i> , <i>Recent Advances in Convolutional Neural Networks</i> , arXiv: 1512.07108v1 [cs.CV] (December 22, 2015) (“Gu”)
1015	Dominik Scherer <i>et al.</i> , <i>Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition</i> , 20 th International Conference on Artificial Neural Networks (ICANN), Thessaloniki, Greece (2010) (“Scherer”)
1016	U.S. Patent No. 9,886,948 to Garimella <i>et al.</i> (“Garimella”)
1017	Dingjun Yu <i>et al.</i> , <i>Mixed Pooling for Convolutional Neural Networks</i> , The 9 th International Conference on Rough Sets and Knowledge Technology (2014) (“Yu”)

Exhibit	Description
1018	Alex Krizhevsky <i>et al.</i> , <i>ImageNet Classification with Deep Convolutional Neural Networks</i> , Advances in Neural Information Processing Systems (2012) (“Krizhevsky”)
1019	Thomas L. Floyd, <i>Digital Fundamentals</i> , 11 th ed., Pearson Education Ltd. 2015, excerpts (“Floyd”)
1020	<i>Intentionally left blank</i>
1021	Yoshua Bengio <i>et al.</i> , <i>Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation</i> , arXiv: 1308.3432v1 [cs.LG] (August 15, 2013) (“Bengio 1”)
1022	Emmanuel Bengio <i>et al.</i> , <i>Conditional Computation in Neural Networks for Faster Models</i> , arXiv: 1511.06297v1 [cs.LG] (November 19, 2015) (“Bengio 2”)
1023	Vincent Vanhoucke <i>et al.</i> , <i>Improving the speed of neural networks on CPUs</i> , Deep Learning and Unsupervised Feature Learning Workshop, NIPS (2011) (“Vanhoucke”)
1024	Patrick Judd <i>et al.</i> , <i>Reduced-Precision Strategies for Bounded Memory in Deep Neural Nets</i> , International Conference on Learning Representations 2016 (2016) (“Judd”)
1025	Microsoft Computer Dictionary, 3 rd ed., 1997, excerpts (“Computer Dictionary”)
1026	Song Han <i>et al.</i> , <i>EIE: Efficient Inference Engine on Compressed Deep Neural Network</i> , arXiv:1602.01528v2 [cs.CV] (May 3, 2016) (“Han”)
1027	Vinay K. Chippa <i>et al.</i> , <i>Analysis and Characterization of Inherent Application Resilience for Approximate Computing</i> , DAC ’13: Proceedings of the 50 th Annual Design Automation Conference (2013) (“Chippa”)
1028	Pakorn Watanachaturaporn <i>et al.</i> , <i>Hyperspectral Image Classification Using Support Vector Machines: A Comparison with Decision Tree and Neural Network Classifiers</i> , ASPRS 2005 Annual Conference (2005) (“Watanachaturaporn”)
1029	Wei Wen <i>et al.</i> , <i>An EDA Framework for Large Scale Hybrid Neuromorphic Computing Systems</i> , DAC ’15: Proceedings of the 52 nd Annual Design Automation Conference (2015) (“Wen”)
1030	<i>Intentionally left blank</i>
1031	Yann LeCun <i>et al.</i> , <i>Deep learning</i> , Nature Vol. 521 (May 28, 2015) (“LeCun 2”)

Exhibit	Description
1032	Ahmed AK Tahir, <i>Integrating artificial neural network and classical methods for unsupervised classification of optical remote sensing data</i> , EURASIP Journal on Advances in Signal Processing (2012) (“Tahir”)

21. I have considered these materials from the viewpoint of a POSITA as of the priority date of the '831 Patent. For the purposes of this declaration, I have been asked to assume that the priority date of the '831 Patent is September 26, 2016. I note that my opinions provided in this declaration are made from the perspective of a POSITA as of this priority date of the '831 Patent, unless expressly stated otherwise. To the extent that I use any verb tense in this declaration that is present tense (*e.g.*, “a POSITA would understand” instead of “a POSITA would have understood”), such verb tense should be understood to be my opinion as of the '831 Patent’s priority date (again, unless expressly stated otherwise). I merely use the present verb tense for ease of reading.

III. LEGAL FRAMEWORK

22. I am a technical expert and do not offer any legal opinions. However, counsel has informed me as to certain legal principles regarding patentability and related matters under United States patent law, which I have applied in performing my analysis and arriving at my technical opinions in this matter.

23. I have been informed that the Patent Trial and Appeal Board (“PTAB”) now applies the *Phillips* standard regardless of whether a patent has expired. I have

been informed that under the *Phillips* standard, claim terms are to be given the meaning they would have to a person having ordinary skill in the art at the time of the invention, taking into consideration the patent, its file history, and, secondarily, any applicable extrinsic evidence (*e.g.*, dictionary definitions). I understand that, for purposes of this proceeding, Petitioner applies the ordinary and customary meaning of the claim terms as understood by a POSITA. I have applied the plain meaning of the claim language as instructed by *Phillips*.

24. I have also been informed that a person cannot obtain a patent on an invention if the differences between the invention and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art. I have been informed that a conclusion of obviousness may be founded upon more than a single item of prior art. I have been further informed that obviousness is determined by evaluating the following factors: (1) the scope and content of the prior art, (2) the differences between the prior art and the claim at issue, (3) the level of ordinary skill in the pertinent art, and (4) secondary considerations of non-obviousness. In addition, the obviousness inquiry should not be done in hindsight. Instead, the obviousness inquiry should be done through the eyes of a person of ordinary skill in the relevant art at the time the patent was filed.

25. In considering whether certain prior art renders a particular patent claim obvious, counsel has informed me that I can consider the scope and content of the prior art, including the fact that one of skill in the art would regularly look to the disclosures in patents, trade publications, journal articles, industry standards, product literature and documentation, texts describing competitive technologies, requests for comment published by standard setting organizations, and materials from industry conferences, as examples. I have been informed that for a prior art reference to be proper for use in an obviousness analysis, the reference must be “analogous art” to the claimed invention. I have been informed that a reference is analogous art to the claimed invention if: (1) the reference is from the same field of endeavor as the claimed invention (even if it addresses a different problem); or (2) the reference is reasonably pertinent to the problem faced by the inventor (even if it is not in the same field of endeavor as the claimed invention). In order for a reference to be “reasonably pertinent” to the problem, it must logically have commended itself to an inventor’s attention in considering his problem. In determining whether a reference is reasonably pertinent, one should consider the problem faced by the inventor, as reflected either explicitly or implicitly, in the specification. I believe that all of the references that my opinions in this IPR are based upon are well within the range of references a person of ordinary skill in the art would consult to address the type of problems described in the Challenged Claims.

26. I have been informed that, in order to establish that a claimed invention was obvious based on a combination of prior art elements, a clear articulation of the reason(s) why a claimed invention would have been obvious must be provided. Specifically, I am informed that a combination of multiple items of prior art renders a patent claim obvious when there was an apparent reason for one of ordinary skill in the art, at the time of the invention, to combine the prior art, which can include, but is not limited to, any of the following rationales: (A) combining prior art methods according to known methods to yield predictable results; (B) substituting one known element for another to obtain predictable results; (C) using a known technique to improve a similar device in the same way; (D) applying a known technique to a known device ready for improvement to yield predictable results; (E) trying a finite number of identified, predictable potential solutions, with a reasonable expectation of success; (F) identifying that known work in one field of endeavor may prompt variations of it for use in either the same field or a different one based on design incentives or other market forces if the variations are predictable to one of ordinary skill in the art; or (G) identifying an explicit teaching, suggestion, or motivation in the prior art that would have led one of ordinary skill to modify the prior art reference or to combine the prior art references to arrive at the claimed invention.

27. I am informed that the existence of an explicit teaching, suggestion, or motivation to combine known elements of the prior art is a sufficient, but not a

necessary, condition to a finding of obviousness. This so-called “teaching suggestion-motivation” test is not the exclusive test and is not to be applied rigidly in an obviousness analysis. In determining whether the subject matter of a patent claim is obvious, neither the particular motivation nor the avowed purpose of the patentee controls. Instead, the important consideration is the objective reach of the claim. In other words, if the claim extends to what is obvious, then the claim is invalid. I am further informed that the obviousness analysis often necessitates consideration of the interrelated teachings of multiple patents, the effects of demands known to the technological community or present in the marketplace, and the background knowledge possessed by a person having ordinary skill in the art. All of these issues may be considered to determine whether there was an apparent reason to combine the known elements in the fashion claimed by the patent.

28. I also am informed that in conducting an obviousness analysis, a precise teaching directed to the specific subject matter of the Challenged Claim need not be sought out because it is appropriate to take account of the inferences and creative steps that a person of ordinary skill in the art would employ. The prior art considered can be directed to any need or problem known in the field of endeavor at the time of invention and can provide a reason for combining the elements of the prior art in the manner claimed. In other words, the prior art need not be directed towards solving the same specific problem as the problem addressed by the patent. Further, the

individual prior art references themselves need not all be directed towards solving the same problem. I am informed that common sense is important and should be considered. Common sense teaches that familiar items may have obvious uses beyond their primary purposes.

29. I also am informed that the fact that a particular combination of prior art elements was “obvious to try” may indicate that the combination was obvious even if no one attempted the combination. If the combination was obvious to try (regardless of whether it was actually tried) or leads to anticipated success, then it is likely the result of ordinary skill and common sense rather than innovation. I am further informed that in many fields it may be that there is little discussion of obvious techniques or combinations, and it often may be the case that market demand, rather than scientific literature or knowledge, will drive the design of an invention. I am informed that an invention that is a combination of prior art must do more than yield predictable results to be non-obvious.

30. I am informed that for a patent claim to be obvious, the claim must be obvious to a person of ordinary skill in the art at the time of the invention. I am informed that the factors to consider in determining the level of ordinary skill in the art include (1) the educational level and experience of people working in the field at the time the invention was made, (2) the types of problems faced in the art and the

solutions found to those problems, and (3) the sophistication of the technology in the field.

31. I am informed that it is improper to combine references where the references teach away from their combination. I am informed that a reference may be said to teach away when a person of ordinary skill in the relevant art, upon reading the reference, would be discouraged from following the path set out in the reference, or would be led in a direction divergent from the path that was taken by the patent applicant. In general, a reference will teach away if it suggests that the line of development flowing from the reference's disclosure is unlikely to be productive of the result sought by the patentee. I am informed that a reference teaches away, for example, if (1) the combination would produce a seemingly inoperative device, or (2) the references leave the impression that the product would not have the property sought by the patentee. I also am informed, however, that a reference does not teach away if it merely expresses a general preference for an alternative invention but does not criticize, discredit, or otherwise discourage investigation into the invention claimed.

32. I am informed that even if a prima facie case of obviousness is established, the final determination of obviousness must also consider "secondary considerations" if presented. In most instances, the patentee raises these secondary considerations of non-obviousness. In that context, the patentee argues an invention

would not have been obvious in view of these considerations, which include: (a) commercial success of a product due to the merits of the claimed invention; (b) a long-felt, but unsatisfied need for the invention; (c) failure of others to find the solution provided by the claimed invention; (d) deliberate copying of the invention by others; (e) unexpected results achieved by the invention; (f) praise of the invention by others skilled in the art; (g) lack of independent simultaneous invention within a comparatively short space of time; (h) teaching away from the invention in the prior art.

33. I am further informed that secondary considerations evidence is only relevant if the offering party establishes a connection, or nexus, between the evidence and the claimed invention. The nexus cannot be based on prior art features. The establishment of a nexus is a question of fact. While I understand that the Patent Owner here has not offered any secondary considerations at this time, I will supplement my opinions in the event that the Patent Owner raises secondary considerations during the course of this proceeding.

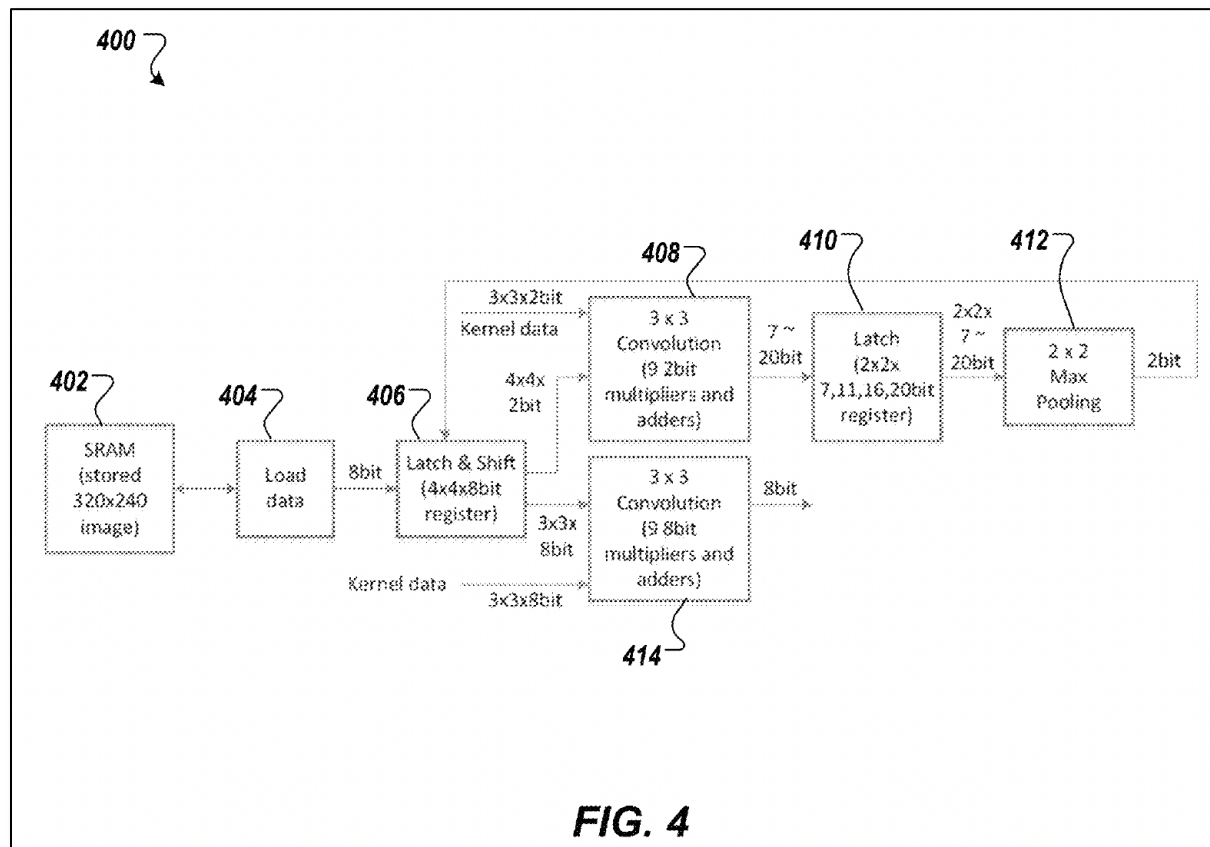
IV. OPINION

A. The '831 Patent

34. The '831 Patent generally describes techniques for “reducing the amount of total computation in convolutional neural networks (CNNs) without affecting the output result or classification accuracy.” *'831 Patent* (EX1001),

Abstract, 1:33-36. The claimed techniques generally recite “performing convolution first with a small number of most significant bits (MSBs), computing the pooling, and if the maximum value is known, performing full-precision convolution on the entry that is the maximum value.” *Id.*, 3:32-37, claim 1. This approximate convolution first, full precision convolution second, technique is depicted below in

Figure 4:

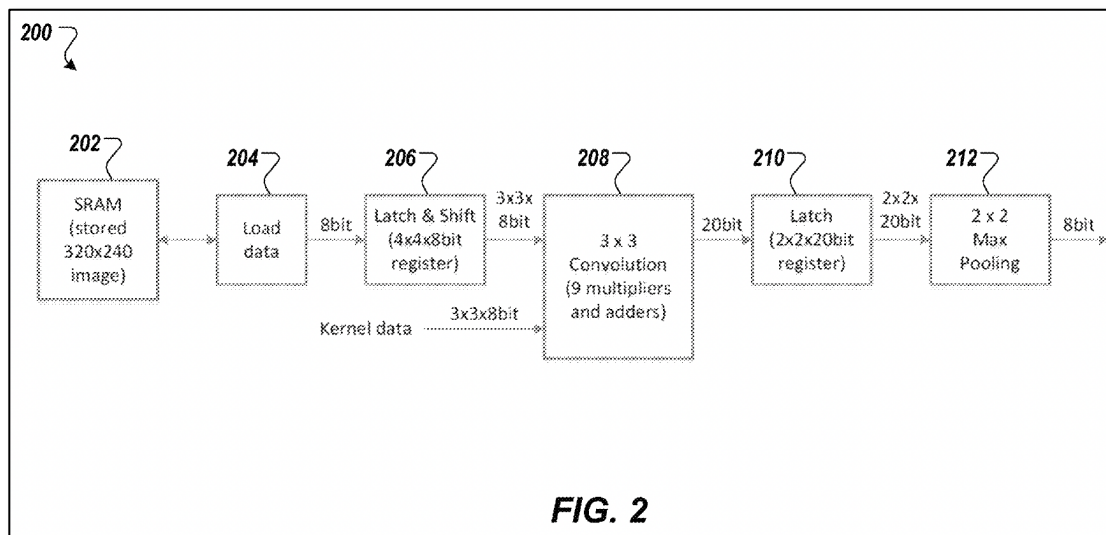


Id., Fig. 4.

35. As shown in Figure 4, “[u]sing the two most significant bits of image data and convolution kernel, approximate multiply-and-add operations are performed (408).” *Id.*, 7:7-23. “The intermediate approximate results are stored into

registers (410), and 2x2 max-pooling is performed on the approximate 2-bit multiply-and-add results (412).” *Id.* “If a maximum is found, the 3x3 convolution may be performed and full precision convolution may be performed on the maximum-exhibiting data set (414).” *Id.* Otherwise, “[i]f a maximum is not found, then the next group of (e.g., higher precision) bits may be retrieved and used to compute results (408) in the next iteration, with any suitable number of iterations performed until a maximum is found.” *Id.*

36. According to the '831 Patent, this contrasts the “conventional convolution and pooling scheme” shown in flow diagram 200 below:



Id., Fig. 2. This exemplary conventional CNN “employs 3x3 convolution, 2x2 pooling, and 8-bit precision.” *Id.*, 5:65-6:1. Here, “an image (e.g., 320x240) may be stored (202) in SRAM[,]” and loaded 204 into registers 206 for processing. *Id.*, 6:12:25. “Convolution (208) may be performed on the data in the registers, e.g., 3x3

convolution based on 3x3 kernel data (e.g., weighting).” *Id.* “The convolution results may be stored in 2x2 bit registers (210), and 2x2 max pooling (212) may be performed to determine the maximum value for the particular layer of analysis before proceeding to the next layer.” *Id.*

37. These conventional methods performed full precision convolutions on every dataset, leading to increase computational and energy expenditure. *Id.*, 3:10-24. The goal of the claimed invention is to purportedly “perform the more fine-grain, full precision, (e.g., more accurate computations on the narrowed-down set of data that exhibited the maximum” to reduce the number of redundant full-precision computations and energy used in the claimed CNN processes. *Id.*, 4:1-4, 9:61-64.

B. Summary of the ’831 Patent’s Prosecution History

38. The ’831 Patent claims its earliest priority to provisional application 62/399,753, filed on September 26, 2016. For purposes of this Petition only, Petitioner applies September 26, 2016, as the priority date for the Challenged Claims.

39. The application received a first action Notice of Allowance after an Examiner’s Amendment to add a missing comma in the pending claims. *’831 File History*, (EX1002), 136-150.

C. Claim Construction

40. I have been informed by counsel that the first step in an unpatentability analysis involves construing the claims, as necessary, to determine their scope. Second, the construed claim language is then compared to the disclosures of the prior art. I am informed that claims are generally given their ordinary and customary meaning as understood by one of ordinary skill in the art at the time of the invention in light of the patent specification and prosecution history, i.e., the intrinsic record. For purposes of my analysis, I have been asked to apply the plain and ordinary meaning as defined in this paragraph.

D. Level of a Person of Ordinary Skill in the Art

41. I understand that the '831 Patent, the record of proceedings at the Patent Office (which I understand is called the "File History" or "Prosecution History"), and the teachings of the prior art, are evaluated from the perspective of a person of ordinary skill in the art ("POSITA") at the time the claimed invention was made. I understand that the factors considered in determining the ordinary level of skill in the art include: (i) the levels of education and experience of persons working in the field; (ii) the types of problems encountered in the field; and (iii) the sophistication of the technology. I may also consider, if available, prior art solutions to the problems encountered in the art and the rapidity with which innovations are made in the relevant art.

42. I understand that a person of ordinary skill in the art is not a specific real individual, but rather a hypothetical individual having the qualities reflected by these factors. This hypothetical person has knowledge of all prior art in the relevant field as if it were arranged on a workshop wall and takes from each reference what it would teach to a person having the skills of a POSITA.

43. Based on my review and analysis of the '831 Patent, the prior art cited herein, and the ordinary skill factors described in this section, a POSITA in the field of the '831 Patent at the time of the earliest possible priority date (September 26, 2016) would have had a bachelor's degree in electrical and/or computer engineering, computer science, or the equivalent, and 2-3 years of work/research experience in convolutional neural networks, artificial neural networks, deep learning, or deep neural networks. Additional education (e.g., MS degree) might substitute for less experience and vice versa.

44. As of September 26, 2016, I had more than ordinary skill in the art and meet at least these minimum qualifications to be a person having ordinary skill in the art as of the time of the claimed invention of the '831 Patent. I am familiar with the skills and knowledge possessed by those I would have considered to be of ordinary skill in the art as of that date. *See above*, Educational Background and Qualifications.

45. My opinions provided in this declaration would not change in view of minor modifications to this level of ordinary skill.

V. BACKGROUND OF THE TECHNOLOGY

46. The features described and claimed in the '831 Patent were routine as of September 26, 2016. In the following paragraphs, I briefly summarize the technology background for a few key features relevant to the opinions set forth herein.

A. Convolutional Neural Networks

47. Convolutional neural networks (CNNs) were well known before the priority date of the '831 Patent. *LeCun 1* (EX1013), 1 (describing, in 1998, a CNN for character recognition on pixel images). In general, CNNs are a type of artificial neural network used for processing data such as in image classification and object recognition.

48. “The basic CNN architecture typically consists of three types of layers, namely convolutional layer, pooling layer and fully-connected layer.” *Gu* (EX1014), 1. The feature extraction phase of the CNN comprises the convolutional and pooling layers. The convolutional layer applies a kernel (i.e., a filter) to the input image, the kernel being a small matrix (e.g., “3x3,” “4x4,” “5x5,” as well as other potential sizes) that slides across the image and performs computations, namely multiplication and addition, with the corresponding image pixels. *LeCun 1*, 6, 13, 35-36. The

computations create dot product maps of the image, also called ‘feature maps,’ which help to reveal edges, curves, textures, or other aspects of the image. *LeCun 1*, 6.

49. The pooling layer samples the feature map by reducing each 3x3 kernel to a single value. *Scherer* (EX1015), 1, 4 (“[t]he pooling layers reduce the size of the input by summarizing neurons from a small spatial neighborhood.”). Thus, the result of pooling is a feature map of reduced dimensionality. *LeCun’s* seminal paper on CNNs describes the creation of a feature map via the “extract[ion of] elementary visual features such as oriented edges, end-points, corners (or similar features in other signals such as speech spectrograms)[,]” followed by pooling (“[t]hese features are then combined by the subsequent layers in order to detect higher-order features.”). *LeCun 1*, 6. There may be multiple rounds (*i.e.*, iterations) of convolution and pooling. *LeCun 2* (EX1031), 4 (describing the typical CNN architecture as having “[a] few stages” of the convolution and pooling layers, where “[t]wo or three stages” of convolution and pooling is performed).

50. The classification phase of the CNN comprises the fully-connected or ‘dense’ layer(s), which combine the features extracted from the convolutional and pooling layers to make a final prediction (*i.e.*, output) of what appears in the input image. *Gu*, 2 (“[The fully-connected layers] take all neurons in the previous layer

and connect them to every single neuron of current layer. ... The outputs of the last fully-connected layer will be fed to an output layer.”)

B. Max Pooling in CNNs

51. The pooling layer reduces the resolution of the feature maps following a convolution layer. *Gu*, 2 (“The pooling layer aims to achieve spatial invariance by reducing the resolution of the feature maps. It is usually placed between two convolutional layers.”). Pooling may be one of several types, but typically are either average pooling or max pooling. *Gu*, 2 (“The typical pooling operations are average pooling [10] and max pooling [11–13].”; *Scherer*, 4 (“[the pooling layers] combine the input from a small $n \times n$ patch of units[.]”). Max pooling was a well-known type of CNN pooling before the priority date of the ’831 Patent. *Id.*, 2 (showing that by 2010 “[m]ore recently, the subsampling operation in CNNs has been replaced with a max pooling operation”).

52. In max pooling, “only the maximum value within the receptive field is propagated to the next layer.” *Id.*; *Garimella* (EX1016), 1:27-30 (“Some NNs, such as convolutional neural networks, use a technique referred to as ‘max pooling’ in which multiple values are generated at a given layer, and the maximum values are forwarded to the next layer.”). As discussed above, the maximum value represents the highest-order feature within each kernel. *See* §V.A. Average pooling, by comparison, uses the average of the values in each kernel. *Scherer*, 1 (“Most earlier

models perform a subsampling operation, where the average over all input values is propagated to the next layer.”)

53. While some groups have proposed combining average pooling and max pooling techniques, the mechanisms of each are well understood. *See generally, Yu* (EX1017). For example, it is known that max pooling “only considers the maximum element and ignores the others in the pooling region.” *Yu*, 5-6. This is true even when two or more convolutional results in a pooling region exhibit the maximum – the CNN chooses a single one of the maximum values to be passed on to the next layer. *Yu*, 5-6, Fig. 2.

C. Rectified Linear Unit in CNNs

54. A Rectified Linear Unit (ReLU) was known to be “one of the most notable non-saturated activation functions[,]” where activation functions serve to “significantly improve[] the performance of a CNN for a certain task.” *Gu*, 3. Particularly, ReLU is applied after each convolutional and fully-connected layer as a means to increase the efficiency and speed of learning. *Krizhevsky* (EX1018), 3 (“Deep convolutional neural networks with ReLUs train several times faster than their equivalents with tanh units.”). The feature map created in the convolutional layer may include both positive and negative values. ReLU suppresses, or zeroes out, all the negative values while leaving all the positive values unchanged. *Gu*, 3 (“ReLU is a piecewise linear function which prunes the negative part to zero and

retains the positive part.”). The result is a feature map with only useful activations (positive values), which more clearly highlights the important features of the input image and thus the values from which the max pooling layer will select the maximum.

D. Most Significant Bits

55. It was also known before the priority date of the '831 Patent that convolutions may use most significant bits (MSBs). MSBs are the leftmost bits in a binary representation of data. *Floyd* (EX1019), 12 (“The left-most bit is the MSB[.]”) This bit is ‘most significant’ in that it contributes the largest weight to the value of the number. *Id.* (“The weight or value of a bit increases from right to left in a binary number.”). *Floyd* provides an example of a binary number 1101101, the leftmost 1 is the MSB. *Id.*, 13. This binary number is equivalent to the decimal number 109, with the MSB being equivalent to 64 and thus contributing well over half (59%) of the number’s weight. *Id.*

56. MSBs are thought of as low-precision values because their heavy weight provides only a coarse overall representation of a piece of data. In contrast to MSBs, the least significant bits (LSBs) are the rightmost bits in a binary number and have a weight of $2^0 = 1$. *Floyd*, 12. Thus, while LSBs contribute the least weight to the value of the number, they may determine the finer details of the data if necessary. In other words, a CNN using LSBs may provide more accurate output.

The benefit to performing convolutions on less than the full dataset, e.g., MSBs only, is that it reduces the total computations in a CNN and thus provides a more efficient process in terms of both time and power resources. *See* §V.E below.

E. Improving Efficiency in Deep Learning

57. Within the field of deep learning, the concept of accelerating computational processing while maintaining prediction accuracy was well-known prior to the '831 Patent. Accelerating computational processing in CNNs may be achieved, for instance, by reducing the total computations required, which thereby increases the overall speed of the computations.

58. As one example, *Bengio 1* discusses conditional computation, “by which only a small subset of the model parameters are ‘activated’ (and need to be visited) for any particular example, thereby greatly reducing the number of computations needed per example.” *Bengio 1* (EX1021), 2, *see also id.*, 1 (“The resulting sparsity [of the activated units] can be potentially [] exploited to greatly reduce the computational cost of large deep networks[.]”). *Bengio 2* also explored the problem that “the evaluation and training of deep models can be time-consuming and computationally expensive.” *Bengio 2* (EX1022), 1. Reinforcement learning was introduced in the conditional computation, which included a reinforcement signal (“a policy gradient algorithm for learning policies that optimize this loss function”) and regularization mechanism (“that encourages diversification of [node] dropout”).

Bengio 2, 1, 7. The resulting method achieved a deep learning model with faster computation speed while still preserving prediction accuracy. *Bengio* 2, 8 (“Our experiments show that it is possible to train such models at the same levels of accuracy as their standard counterparts. Additionally, it seems possible to execute these similarly accurate models faster due to their sparsity.”).

59. As another example, Vanhoucke et al. discusses how to improve the speed of neural networks on CPUs in the speech recognition field. “[T]he use of large, deep neural networks with tens of millions of parameters ... can represent a challenging computational burden, even for modern CPUs.” *Vanhoucke* (EX1023), 1. The paper introduces different techniques to reduce this computational cost on CPUs. *Id.*, 1 (“We use speech recognition as an example task, and show that a real-time hybrid hidden Markov model / neural network (HMM/NN) large vocabulary system can be built with a 10x speedup over an unoptimized baseline and a 4x speedup over an aggressively optimized floating-point baseline at no cost in accuracy.”), 8 (“We showed that [by] leveraging these faster [fixed-point SIMD (Single Instruction, Multiple Data)] instructions, a real-time speech recognizer can be built using a very large hybrid network at no cost in accuracy”).

F. Adaptive Precision

60. Adaptive precision was a known technique for improving computational efficiency in the field of computer vision before the priority date of

the '831 Patent. For example, Judd et al. describe methods for “[t]uning precision per layer” of CNNs to “enable energy and performance improvements.” *Judd*, (EX1024), Abstract. *Judd* analyzed “conventional” implementations of CNNs using 32-bit floating point representations of the data versus CNNs with reduced precision. *Id.* Here, the authors varied the bit-width of the weights and data input by adjusting the number of integer and fractional bits. *Id.*, 3, Fig. 3. The authors recognized that this type of adaptive precision, using varying bit widths per CNN layer, could lead to “potential benefits includ[ing] energy reduction, higher performance and the possibility of supporting larger networks.” *Id.*, 9.

VI. OPINION: GROUND 1

A. Obviousness over Ujiie and Moons

1. Overview of Ujiie

61. I have reviewed the Declaration of Ms. June Munford regarding her methodology in identifying and locating *Ujiie*. *Munford Dec.* (EX1009), ¶¶11-26. As a person with more than ordinary skill in the art, I agree with Ms. Munford’s methodology in identifying and locating *Ujiie*.

62. In my over 30 years in the fields of computer vision and machine learning, I have found that open access databases storing conference papers are freely accessible to at least attendees of the conference. This repository is crucial for providing access to attendees both during and following the conference. These publications are similarly crucial for research entities (e.g., academic labs) to

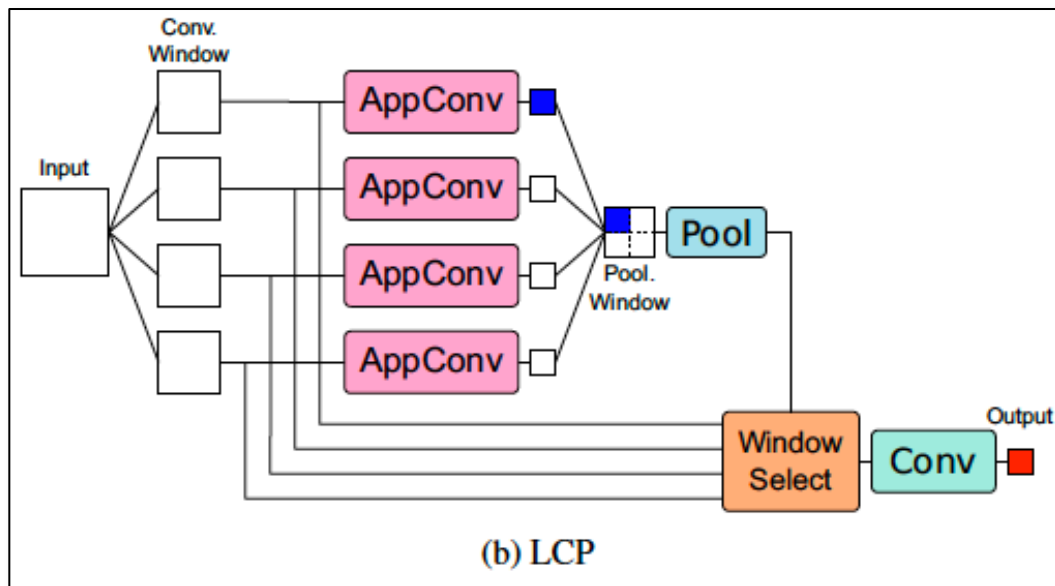
publicly display progress in certain areas so that the entity can maintain and apply for additional research funding. I have personally searched, downloaded, and reviewed many conference publications, both as an attendee and as a non-attendee. I have also personally submitted multiple articles for publications at similar conferences, including IEEE conferences.

63. Specifically, I have been asked to opine on whether *Ujiie* would have been accessible to persons interested and ordinarily skilled in the subject matter. I have also been asked to consider whether, even without knowledge of the work, the *Ujiie* article title is adequately descriptive such that an individual interested in this research area would have located it through ordinary means. I understand *Ujiie* is a publication “made available no later than the last day” of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. *MacPherson Dec.* (EX1010), ¶13 (opining that *Ujiie* was made available at least by July 1, 2016). I have also been made aware of a Declaration from Ms. June Munford that describes how the publication was available online via the CVF database at least as of August 5, 2016. *Munford Dec.*, ¶¶11-26. Ms. Munford describes locating *Ujiie* through CVF’s 2016 webpage, which hosted the CVPR proceedings. *Id.* Namely, Ms. Munford located *Ujiie* by selecting the ‘Embedded Vision’ on the proceedings webpage. *Id.*, ¶¶16-21.

64. I agree with Ms. Munford's methodology, and it is therefore my opinion that a person interested in this subject matter could indeed have accessed the Open Access CVF webpage hosting the 2016 CVPR proceedings and discovered *Ujiie* as of August 5, 2016. CVPR conference proceedings are known to persons interested in computer vision and pattern recognition techniques, including those interested in embedded vision and machine learning (e.g., CNNs). An interested individual, even without prior knowledge of *Ujiie*, would have been able to locate the article through a reasonable search given the description of both the conference and the subcategory of embedded vision. In fact, CVPR is one of the premier conferences in the computer vision and pattern recognition fields. Interested individuals (e.g., POSITAs) working in these areas would be well aware of CVPR and its proceedings.

65. I additionally agree that the attendees of the CVPR conference would indeed be those interested in computer vision and, based on my prior experience with these conferences in combination with Mr. MacPherson's Declaration, would have had access to *Ujiie* at least as of the last date of the conference. *MacPherson Dec.*, ¶13. Furthermore, I have been made aware by Ms. Munford's Declaration that the conference's actual attendance exceeded over 3,000 attendees as well as 50 companies. *Munford Dec.*, ¶23. Accordingly, a large group of interested persons would have had access to, and would have easily found, *Ujiie* at the conference.

66. *Ujii* is generally directed to an “approach[] for realizing energy-efficient CNN[s]” by pursuing “the redundancy in most of the recognition algorithms[.]” *Ujii* (EX1004), 1. To decrease redundancy, *Ujii* teaches “approximated convolution units (AppConv) to predict the feature selection in the pooling layer, and [then] a single regular (and exact) convolution unit (Conv) to perform convolution only on the predicted window.” *Id.*, 4. This process is illustrated in Figure 2b below:



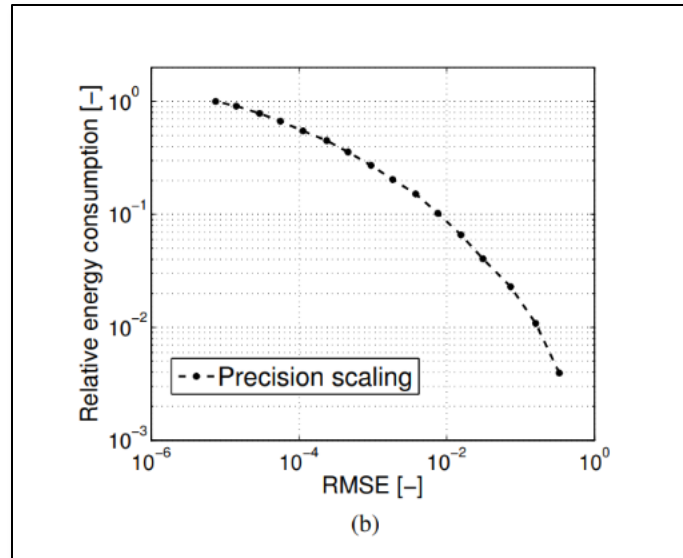
Id., Fig. 2b.

67. *Ujii*’s process, termed “LazyConvPool (LCP),” is “an approximate computational method that combines convolution and max pooling for reducing power consumption of CNN processors.” *Id.*, 3. “The computation flow of LCP is summarized as follows: Step 1 Compute AppConv using the lightweight and approximate computation. Step 2 Operate max pooling to the results of AppConv to

specify the convolution index of which the convolution result is propagated. Step 3 Compute Conv only to the input corresponding to the specified index in Step 2.” *Id.*, 4. Using this process, *Ujiie*’s system “only executes an exact convolution (Conv) against the specified window predicted by the approximate convolution (AppConv). *Id.*, 3. *Ujiie*’s approximation process “reduce[s] the number of multiplications in CNN[,]” thereby “reduc[ing] the energy consumption for image recognition by 11.4%–14.1% while keeping the recognition accuracy.” *Id.*, 6.

2. Overview of Moons

68. *Moons* “proposes methods based on approximate computing to reduce energy consumption in state-of-the-art ConvNet accelerators.” *Moons* (EX1005), 1. *Moons* “quantize[s] existing networks and drastically reduce[s] the number of bits that encode the weights and inputs of each layer” while maintaining accuracy. *Id.* This both “can be used to reduce energy consumption in many neural network applications and even dynamically scale energy consumption depending on the current application.” *Id.* Specifically, *Moons* employs “precision scaling, in which the number of bits used in computations is varied at run-time[.]” *Id.*, 2. “By scaling precision (i.e., dynamically scaling the number of bits that encode the network’s weights and inputs),” power consumption is reduced. *Id.*; see also *id.*, 5, Fig. 1.



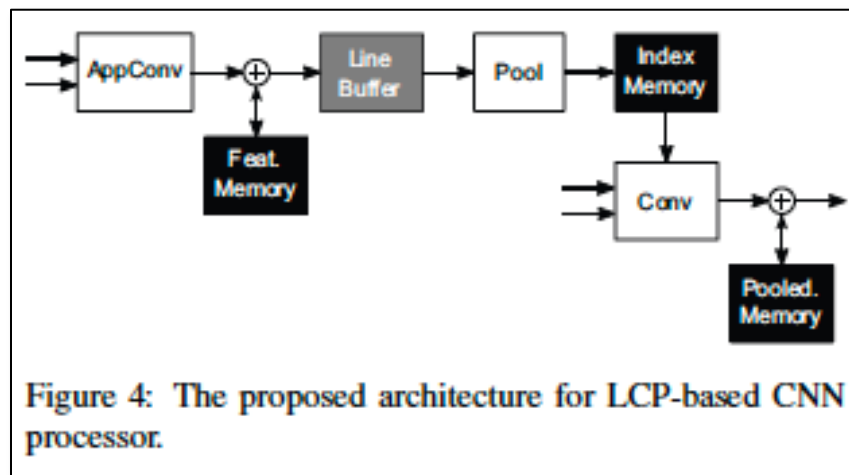
Id., Fig.1b. For example, *Moons* teaches quantizing weights and inputs to “4-10 bits instead of the previously used 16- or 32-bit formats” so long as the “MSB-bit” is “placed at the MSB position.” *Id.*, 5, 6.

3. Claim 1(Pre): “One or more non-transitory computer-readable storage media storing instructions which, when executed by at least one processor, cause the at least one processor to perform operations comprising:”

69. To the extent the preamble is limiting, *Ujii* teaches a “base architecture of [a] **CNN processor**...[that] employ[s] a parallel architecture with a **shared memory** and **multiple processing elements**. Each processing element, connected to the shared memory, executes convolution, pooling, and activation.” *Ujii*, 4¹. Specifically, *Ujii* teaches a “processing element architecture in which LCP

¹ All emphasis added unless otherwise noted.

[LazyConvPool] is applied to the base [CNN processor] architecture.” *Id.*, 4-5. This proposed architecture is depicted below in Figure 4:



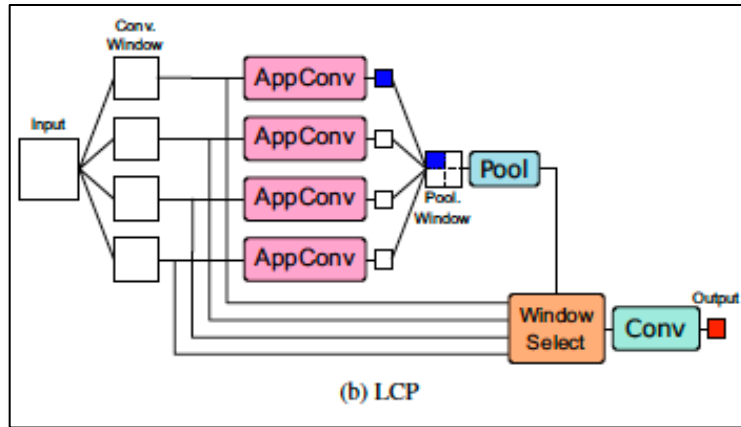
Id., Fig. 4.

70. *Ujiie*’s processing element “executes” convolution, pooling, and activation. *Id.*, 4; *see also id.* 4-5 (“A processing element [] computes AppConv and pooling. The results are temporarily stored in the index memory connected to the pooling circuit and utilized as the signal that control inputs to the Conv circuit. Conv results are accumulated only for the size of the pooled feature map.”). A POSITA would have understood that executing processing steps requires computer-executable instructions to be stored in storage media (e.g., persistent storage such as a hard drive or solid-state drive). *Computer Dictionary* (EX1025), 337 (“object code...*n*. The code, generated by a compiler or an assembler, that was translated from the source code of a program. The term most commonly refers to machine code that can be directly executed by the system’s central processing (CPU), but it can

also be assembly language source code or a variation of machine code. *See also* central processing unit.”), 84 (discussing central processing unit), 443 (discussing source code). Therefore, a POSITA would have understood that *Ujiie*’s non-transitory shared memory stores instructions that when executed by its “CNN processor[,]” performs the functions of the processing elements (e.g., AppConv, LCP, convolution, pooling, and activation) in *Ujiie*’s proposed architecture.

4. Claim 1(a): “in one or more layers of a convolutional neural network (CNN), performing a first iteration that includes computing a value based on a first set of most significant bits (MSBs) for each of a plurality of data sets;”

71. *Ujiie* performs a first iteration in a convolution layer where its “LazyConvPool (LCP)” computing method computes an “approximate convolution (AppConv)” value for each dataset related to its input image. This AppConv calculation satisfies the claimed “computing a value” step (i.e., *in one or more layers of a convolutional neural network (CNN), performing a first iteration that includes computing a value*). *Ujiie*, 1-4. LCP is “a new computation method for CNN’s convolutional layers and pooling layers.” *Id.*, 1. The “LCP framework [has] two main parts: (1) **approximated convolution units (AppConv) to predict the feature selection in the pooling layer**, and (2) a single regular (and exact) convolution unit (Conv) to perform convolution only on the predicted window.” *Id.*, 4. Figure 2b illustrates the LCP-based convolution and pooling operations:



Id., Fig. 2b.

72. In more detail, “[b]uffered inputs are supplied to AppConv and Conv with weights used for convolution.” *Id.*, 4. AppConv outputs “an **approximate computation unit** of reduced computation complexity.” *Id.* The input image, convolution window datasets, and AppConv values are shown above in Figure 2b. *Id.*, Fig. 2b; *see also id.*, Tables 1 and 2 (depicting 28x28 or 32x32 image convolved by 5x5 convolution windows).

Layer name	Parameter
Input image	size: 28x28, channel: 1
Convolution	kernel: 5x5, channel: 20
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 50
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 500
ReLU	
Fully connected	channel: 10
Softmax	

Table 1: Network parameters of CNN for MNIST dataset.

Layer name	Parameter
Input image	size: 32x32, channel: 3
Convolution	kernel: 5x5, channel: 64, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 128, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 256, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 512
ReLU	
Fully connected	channel: 10
Softmax	

Table 2: Network parameters of CNN for CIFAR-10 dataset.

Id., Tables 1 (top) and 2 (bottom). To calculate these approximate convolution (AppConv) values, *Ujiie* implements an “approximate computation called Sign Connect[.]” *Id.*, 4.

73. Sign Connect replaces “general multipliers” “by the multiplications of the sign of weights, which can be typically realized by multiplexers.” *Id.* In other words, as given in equations (8) and (9) of *Ujiie*, if the sign of the weight is > 0 , the weight is binarized to +1, whereas if the weight is < 0 , the weight is binarized to -1, as shown below:

$$\begin{cases} P(W(k, l) = 1) = 1 \\ P(W(k, l) = 0) = 0 \end{cases}, \quad \begin{cases} P(W(k, l) = -1) = 1 \\ P(W(k, l) = 0) = 0 \end{cases}.$$

Ujiie, Equations (8) and (9). Binarizing the weight to only consider the sign reduces computational cost by lowering power consumption. *Ujiie*, 4. *See also, Courbariaux* (EX1011), 1 (“During the forward pass, BNNs [Binarized Neural Networks] drastically reduce memory size and access, and replace most arithmetic operations with bit-wise operations, which is expected to substantially improve power-efficiency.”) This simplifies the multiplication computations occurring during convolutions. *Ujiie*, 4 (“With Sign Connect, multiplication could be drastically simplified, and the total performance becomes deterministic.”).

74. *Ujiie’s* AppConv values are computed within a CNN layer (e.g., convolutional layer and/or convolutional and pooling layer). *Ujiie*, Tables 1 and 2 (“Layer name” “Convolution”).

Layer name	Parameter
Input image	size: 28x28, channel: 1
Convolution	kernel: 5x5, channel: 20
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 50
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 500
ReLU	
Fully connected	channel: 10
Softmax	

Table 1: Network parameters of CNN for MNIST dataset.

Layer name	Parameter
Input image	size: 32x32, channel: 3
Convolution	kernel: 5x5, channel: 64, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 128, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 256, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 512
ReLU	
Fully connected	channel: 10
Softmax	

Table 2: Network parameters of CNN for CIFAR-10 dataset.

Id., Tables 1 (top) and 2 (bottom). For a number of reasons, set forth in the next paragraph, a POSITA would understand that *Ujii* performs a “first iteration” computing AppConv values consistent with the ’831 Patent’s description of an “iteration.” *Id.*, Figure 2b.

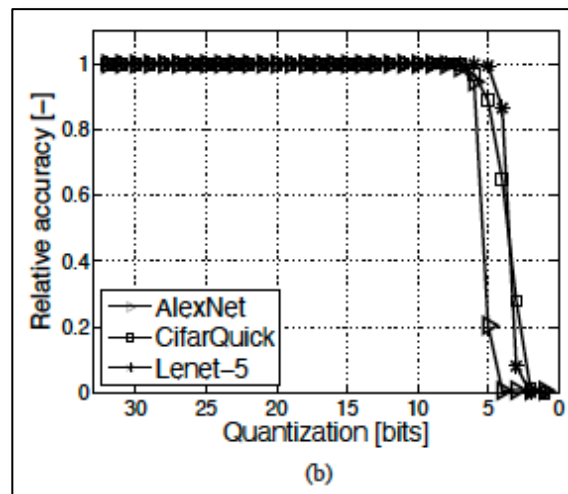
75. First, the ’831 Patent describes “iterations” as rounds of approximate convolutions up to full precision. ’831 *Patent*, 2:6-8 (“the first iteration computes a value that approximates the full precision computation of the value”), 3:58-61 (“At each iteration...a decision may be made whether the next approximate convolution iteration needs to be performed”), 4:4-6 (same), 5:50-51 (“The algorithm may perform any suitable number of iterations, up to the full precision [] convolution”), 7:65-8:2 (the process may **perform another iteration by “performing a convolution** using more bits of data”). Second, *Ujii*’s AppConv is an approximate convolution computation and is performed as a first step. *Ujii*, 4 (“The computation flow of LCP is summarized as follows: **Step 1 Compute AppConv using the lightweight and approximate computation.** **Step 2** Operate max pooling to the results of AppConv to specify the convolution index of which the convolution result is propagated. **Step 3** Compute Conv only to the input corresponding to the specified index in Step 2.” (bold emphases original)). Similar to the ’831 Patent, *Ujii*’s Step 1 AppConv is an “**approximate computation.**” *Ujii*, 4; *see also*, ’831 *Patent*, 3:65-4:4 (“Implementations perform approximate computations of neural networks (e.g.,

based on the MSBs), make a decision based on the approximate results which data set to proceed with, and then (e.g., only) perform the more fine-grain, full precision, (e.g., more accurate) computations on the narrowed-down set of data that exhibited the maximum.”), 2:6-8 (“the first iteration computes a value that **approximates the full precision computation** of the value”). Therefore, *Ujiie*’s system performs a first iteration that includes approximate convolution values (AppConv) for each dataset.

76. While *Ujiie* teaches approximating convolutions using Sign Connect, *Ujiie* does not expressly discuss approximating convolutions by lowering precision “based on a first set of most significant bits[.]” *Moons* teaches approximating convolutions by lowering precision with precision scaling.

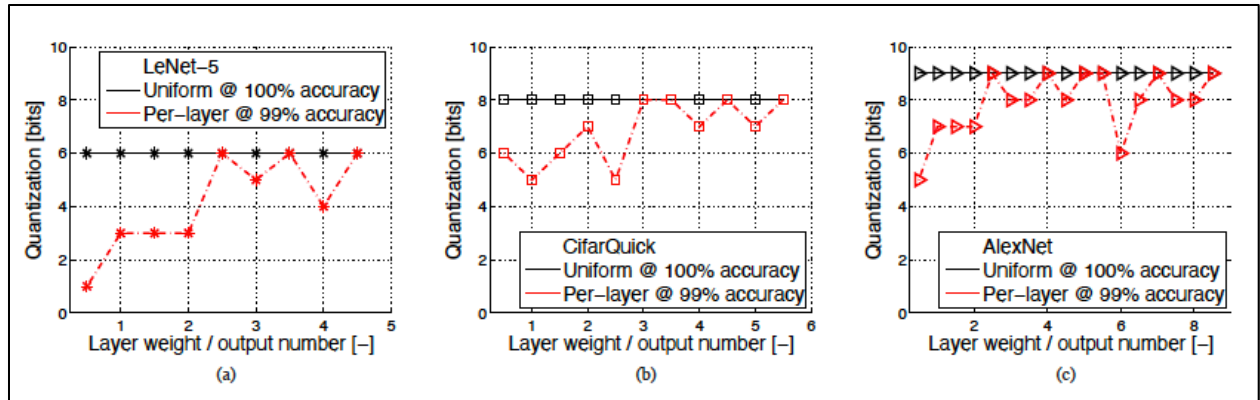
77. To appropriately address the “increase in energy consumption” in CNNs, *Moons* utilizes “precision scaling,” which encompasses “quantiz[ing] existing networks and drastically **reduc[ing] the number of bits that encode the weights and inputs of each layer.**” *Moons*, 1. “Precision scaling is an approximate computing technique allowing major energy savings in digital circuits.” *Id.*, 5. *Moons* teaches that typical “ConvNets run on high precision machines using 32-bit floating point number[s,]” but “such high precision is not always necessary[.]” *Id.*, 3. *Moons* therefore reduces precision by reducing the number of bits used for the inputs (e.g., image data pixel values) and the kernel weights. *Moons* teaches that this bit reduction can be applied either by “[u]niform rescaling: all inputs and weights

are scaled with the same value” or “[p]er-layer rescaling: inputs and weights are scaled on a per-layer basis.” *Id.*, 3. *Moons* goal is to save energy while minimizing accuracy loss. *Id.*, 1 (“we can gain energy...up to 30x without losing classification accuracy and more than 100x at 99% classification accuracy, compared to the commonly used 16-bit fixed point number format”). This accuracy / energy tradeoff is depicted below in Figure 2b:



Id., Fig. 2b (showing accuracy as a function of the bit width of the input and weight for per layer rescaling).

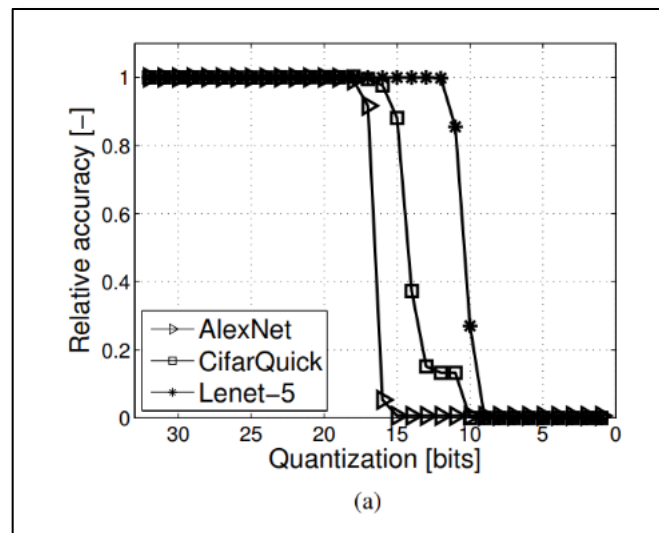
78. *Moons* teaches scaling down precision for three different datasets while maintaining sufficient accuracy:



Id., Figs. 4a-4c. For LeNet, precision is scaled to 1-6 bits; for CifarQuick, precision is scaled to 5-8 bits; and for AlexNet, precision is scaled to 5-9 bits. *Id.*

79. *Moons*, like the '831 Patent, performs approximate computing by implementing precision scaling using the most significant bits of data that encodes weights and inputs of a network. *Moons*, 5 (“**In order to apply precision scaling efficiently** in such a system, architectural changes have to be made. First, the positioning of the sign-bit in the fixed number representation is crucial. **The MSB-bit should always be placed at the MSB position**, otherwise toggling sign bits will lead to high switching activity.”), 2 (“**By scaling precision (i.e., dynamically scaling the number of bits that encode the network’s weights and inputs)**, the switching activity α can be substantially reduced.”); *see also* '831 Patent, 3:53-55 (“Starting from the MSB side, implementations may perform approximate computing group-by-group, and these approximate computing steps are cascaded.”). Both *Moons* and the '831 Patent endeavor to use the least number of bits while achieving an acceptable level of precision in the computation. *Compare, Moons*, 4

(“Figure 2a shows the **relative accuracy** of our three networks as a function of the number of quantization bits. As can be seen in the figure, the relative accuracy stays equal to one for all three networks up to quantization with 19 bits, meaning that **the quantized network reaches the exact same accuracy on its dataset as the original network.**”), with *’831 Patent*, Abstract (“Accordingly, the total number of bits used to perform the convolution is reduced **without affecting the output features or the final classification accuracy.**”)



Moons, Fig. 2a. *Moons* therefore computes approximate convolution values “based on a first set of most significant bits”—i.e., by performing lower precision convolutions with quantized input data and weights containing the most significant bits that allot for sufficient accuracy. *Id.*, 5 (noting in precision scaling, the “MSB-bit should always be placed at the MSB position”).

a. Motivation to Combine Ujiie and Moons

80. A POSITA would have been motivated and found it obvious to compute *Ujiie*'s approximate computations (i.e., AppConv) using *Moons*'s precision scaling approach. Namely, a POSITA would have been motivated to replace *Ujiie*'s Sign Connect approximations (i.e., using a single signed bit for kernel data and full bit-resolution for image data) with *Moons*'s precision scaling quantization techniques (i.e., reduced bit-resolution for both kernel data and image data) to increase system accuracy.

81. *Ujiie*'s Sign Connect computations heavily approximate the kernel value by adjusting all values to simply +1 or -1. *Ujiie*, 2, 4; *see also*, *Courbariaux*, 2 (discussing "sign function" which *Ujiie* explains is equivalent to Sign Connect) and *Ujiie*, 4 ("Sign Connect is equivalent to 'sign function' used in BinaryNet [citing *Courbariaux*].") In other words, *Ujiie* approximates convolutions by using fast logic that cancels out the amplitude of the weight value. For example, weight values that fall well above or well below zero are effectively converted to a much smaller number (i.e. +1 or -1 respectively). This results in a convolution whereby a weight value that would otherwise have a much larger effect on the multiplication with the input data (e.g., multiplying by a weight value of 16) is significantly decreased to just the value of the input data since the input data is "multiplied" by "+1" or "-1" as the case may be. While this process effectively decreases the computational

energy needed to multiply the input data by the weight value, it can significantly decrease the accuracy of the resulting convolutional value used for max pooling. This decrease in accuracy can potentially lead to errors in selecting the most significant input data that should be advanced to a full precision convolution calculation. The point of the AppConv calculation in Step 1 is to identify this most significant input data to be advanced to the full precision convolution calculation, so any errors in the AppConv process that result from inadequate accuracy may lead to advancing the wrong input data to the full precision convolution calculation.

82. Indeed, as *Ujiie* teaches, using Sign Connect approximation introduces errors into the process. For example, *Ujiie* describes how binarization techniques “show[] the performance degradation due to the prediction inaccuracy...for the complex data such as color images.” *Ujiie*, 6. Sign Connect specifically led to a decrease in accuracy for the CIFAR-10 dataset. *Ujiie*, Table 3 (showing an LCP accuracy of 0.860 compared to an LCP + Sign Connect accuracy of 0.849).

Method	MNIST	CIFAR-10
(a) Conventional	0.986	0.860
(b) LCP	0.986	0.860
(c) LCP + TC	0.984	0.825
(d) LCP + SC	0.988	0.849

Table 3: Classification accuracy on datasets.

Id. A POSITA would have been motivated to strike the proper balance by decreasing the inaccuracies caused by Sign Connect, particularly for complex image data using a more precise approximation method while maintaining computational efficiency.

83. *Moons*'s approximations use multi-bit quantization and therefore do not completely remove the weight value's amplitude. Instead, *Moons*'s process performs lower bit-width approximate calculations on all weights and input data, thereby maintaining the most significant portions of the value's amplitude. Indeed, the position of a bit within a binary representation of a value increases in specificity while moving from left (i.e., MSB) to right (i.e., LSB). For example, in an 8-bit binary representation of a value, the left-most bit (i.e., MSB) carries the highest weight and therefore has the most influence on the magnitude of the corresponding decimal number, whereas the right-most bit (i.e., LSB) carries the lowest weight and minimally influences the overall value. Specifically, the value represented by the bit position decreases moving left to right, such that the first bit position (i.e. the MSB) in an 8-bit binary represents 128, the second bit position represents 64, the third bit position represents 32, and so on. Accordingly, use of the MSB in convolutional computations provides greater clarity, and therefore more specificity, in approximate computations than does representing any positive value as +1. *Moons*'s process thus offers higher accuracy than *Ujiie*'s Sign Connect process while still achieving significant compression, i.e., improved computation efficiency.

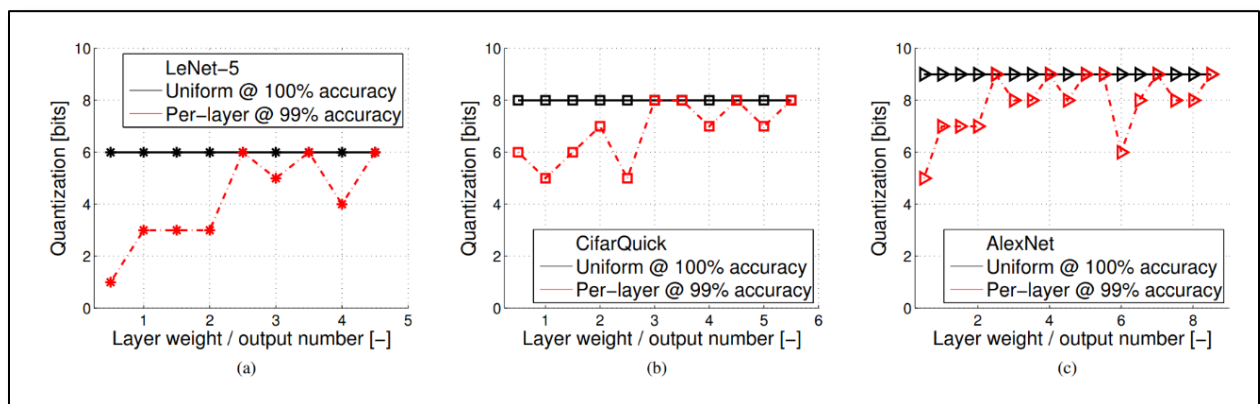
84. Accordingly, a POSITA would have understood that modifying *Ujiie* with *Moons*'s quantization techniques would have enhanced the max pooling step in *Ujiie*'s process. Indeed, as shown above, an 8-bit binary having a "1" at its first bit position establishes a value of at least 128, which is significantly different in convolutional computations than converting this 8-bit value into a +1.

85. Enhancing the accuracy of the AppConv computations allows *Ujiie*'s system to better identify the most significant input data during max pooling. Using the above example, multiplication using a value of 128 to represent the 8-bit better matches what the full precision computation would be as compared to multiplying with only +1. As such, using quantization for *Ujiie*'s approximate convolution step would have decreased the likelihood of error at the max pooling stage, thereby providing more accurate decisions on the most significant data set to perform the computationally expensive full-precision convolution. This modification would therefore enable a POSITA to take advantage of improved precision when appropriate.

86. Moreover, quantization offers more flexibility compared to a limited set of fixed weight values (e.g., +1 and -1). As explained above, the MSBs convey more information regarding the value being represented in binary than does converting any positive value into a "+1". This includes providing an expansive representation of the value within the MSBs (e.g., the first three bits of an 8-bit

binary can represent the values 0, 32, 64, 96, 128, 160, 192, and 224). This example makes clear how a small amount of bits can represent a broader spectrum of values, and therefore offer more flexibility, than the fixed values of +1 or -1.

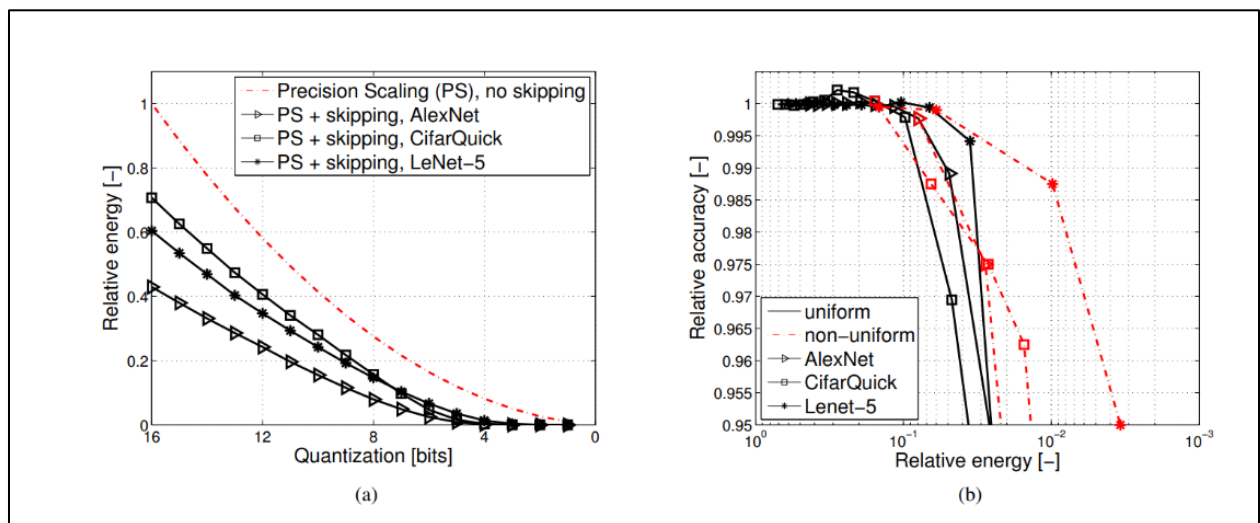
87. Using fixed weights across all layers prevents implementing different precision at different layers that may be more or less tolerant to quantization. *Moons* expressly teaches “per-layer quantization” that finds “an optimal [precision] setting for each layer by exploiting the variations of the layer-specific input and weight distributions.” *Moons*, 4. Replacing fixed weights with quantized representations provides layer-wise tuning, allotting for differing precisions across layers and providing a more robust CNN capable of “control[ling] the energy-accuracy trade-off tightly[.]” *Moons*, 4. *Moons* further teaches that this layer-wise tuning of bit width indeed adjusts as-needed with the layer while still maintaining a certain accuracy. *Moons*, Fig. 4.



Id. In fact, the number of bits necessary to maintain 99% accuracy varied up to 6 bits between layers. *Id.* This variance allows the system to properly balance energy consumption with accuracy.

88. *Ujiie's* system, similar to *Moons*, would have benefited from per-layer quantization at the approximate convolution stage to increase efficiency by using the proper amount of bits to increase the likelihood of identifying a maximum value that accurately identifies the most significant data set, which should be processed in the full resolution convolution. Just as taught in *Moons*, a per-layer quantization system in *Ujiie* would benefit the system by decreasing energy consumption to a level only necessary for accurately approximating the maximum value (i.e., identifying the most significant data set) while using the minimal amount of bits in the approximate convolution computation. For example, the *Ujiie-Moons* system may determine, as taught by *Moons*, that a first layer is typically successful at accurately determining the most significant data set in max pooling using two bits during the approximate convolution step. *Ujiie's* modified system may additionally determine that the second layer may accurately determine the most significant data set in max pooling using three bits. Accordingly, modification of *Ujiie's* system to perform *Moons* per-layer quantization in its approximate convolutions would have provided beneficial flexibility in maintaining necessary accuracy while decreasing energy consumption.

89. While modifying *Ujiie* to perform multiplications of low bit-width may lead to an increase in energy consumption, a POSITA would have recognized that, in some circumstances, the tradeoff for computational energy would be a desirable tradeoff for the enhanced accuracy *Moons's* process realizes. For example, in scenarios involving more complex image analysis (e.g., “color images”), a POSITA would have recognized that improved image analysis using *Moons's* quantization would have benefitted the system striking the proper balance of consuming additional energy to obtain improved image analysis. This modification would have presented a straightforward tradeoff between accuracy and energy consumption based on the number of bits used in the approximation. For example, *Moons* emphasizes this straightforward connection by showing how increased bit width requires the use of more energy. *Moons*, Fig. 6a.



Id., Fig. 6. Regarding accuracy, as taught by *Ujiie*, the Sign Connect process can be solely responsible for a decrease of 1.1% accuracy in some more sophisticated image classification models (e.g., CIFAR). *Ujiie*, Table 3.

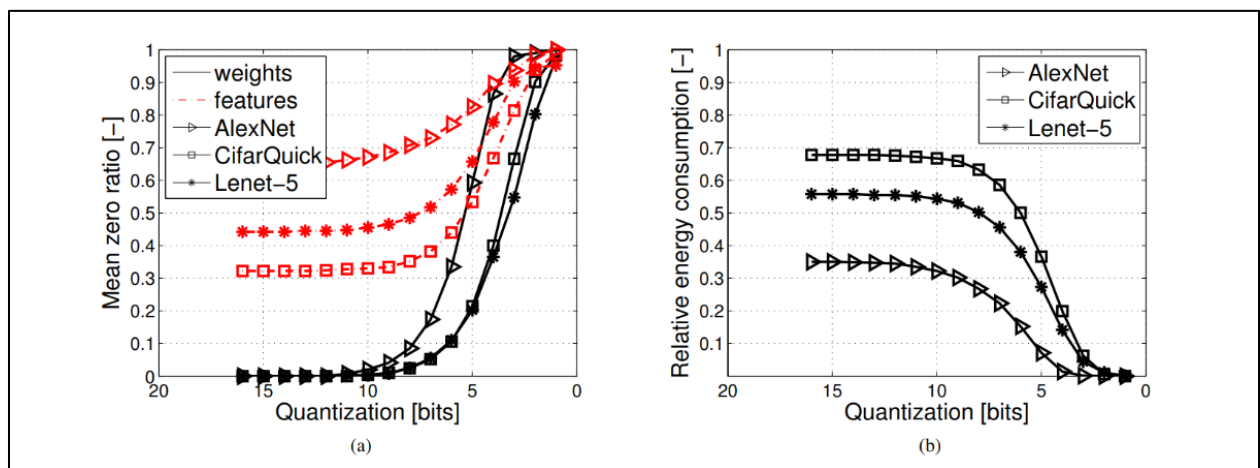
Method	MNIST	CIFAR-10
(a) Conventional	0.986	0.860
(b) LCP	0.986	0.860
(c) LCP + TC	0.984	0.825
(d) LCP + SC	0.988	0.849

Table 3: Classification accuracy on datasets.

Id. While this amount may seem nominal, a POSITA would have understood that CNN image analysis in medical applications, for example, needs to be as precise as possible. These medical applications may often involve those “complex” “color images” taught by *Ujiie*’s system due to Sign Connect. For example, CNNs may be used to diagnose certain skin disorders (e.g., melanoma) which involves analyzing colored images of skin as well as the coloration of melanoma. In these types of exemplary scenarios, a POSITA would have understood that negating the decreased accuracy caused by Sign Connect, such as by using *Moons*’s precision scaling, would have significantly improved the system, even at the cost of a possible increase in energy consumption.

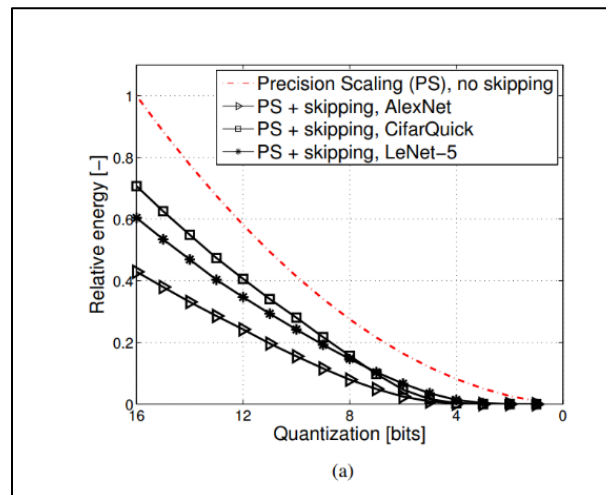
90. Notably, this modification preserves *Ujiie*’s goal of decreasing energy consumption in comparison to a standard CNN approach in which no approximation step is implemented. Indeed, both *Moons* and *Ujiie* describe approximation

techniques that expressly seek to decrease energy consumption, with *Moons*'s approach allowing for increased accuracy, as necessary. *Moons*, Figure 5(b) (showing decreased energy consumption using decreased bit width (i.e., quantization)), 2 ("Approximate computing through precision scaling lowers the active power consumption"), 3 ("To reduce the energy consumption of the ConvNet's computations, our main strategy is to quantize its weights and the inputs to its layers"), 6 ("If per-layer quantization is used...more energy can be gained, while losing less classification accuracy in the process").

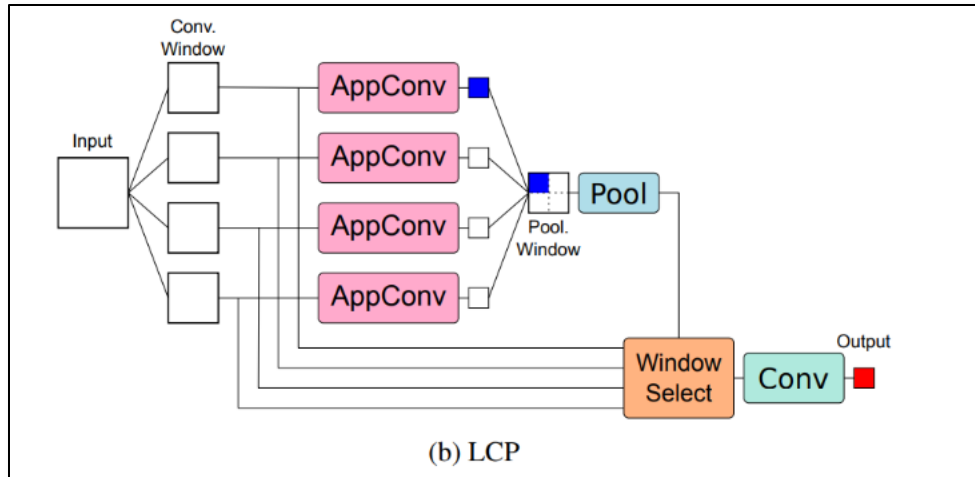


Moons, Fig. 5. Because *Moons*'s approximation still significantly decreases energy consumption compared to typical full precision convolution, a POSITA would have appreciated that *Ujiie*'s modified system would still gain significant improvements in energy consumption, when compared with the traditional convolutional process, while yielding higher accuracy in Step 1.

91. A POSITA would have appreciated a reasonable expectation of success modifying *Ujiie*'s approximate convolution process with *Moons*'s approximated quantization computations because it simply requires replacing one known approximation technique with another to improve *Ujiie*'s process. Both *Ujiie* and *Moons* teach approximate CNN computations for improving energy efficiency. *Compare, Ujiie*, 6 (teaching its LCP algorithm “reduce[s] the energy consumption for image recognition by 11.4%–14.1%”), with *Moons*, Figure 6(a) (showing a significant reduction in relative energy usage as a function of bit width), 6 (teaching that “applying approximate computing” “drastically reduce[s]” “the energy consumption”).

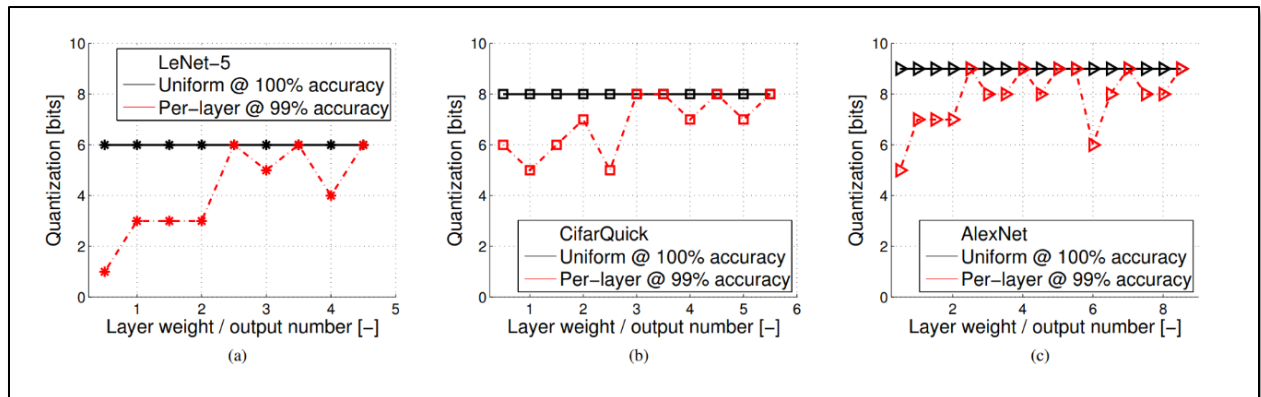


Moons, Fig 6(a). Indeed, *Ujiie*'s approximate CNN computation includes using Sign Connect to approximate a convolutional result to be used in max pooling operations to determine which data set to perform a computationally expensive full precision convolution. *Ujiie*, 3-4, Fig. 2b.



Id., Fig. 2b. *Moons*'s approximation includes precision scaling, whereby a decreased amount of bits are used for convolutional operations, which decreases specificity of the convolution to the necessary amount to maintain a certain level of accuracy.

Moons, 5, Fig. 4.



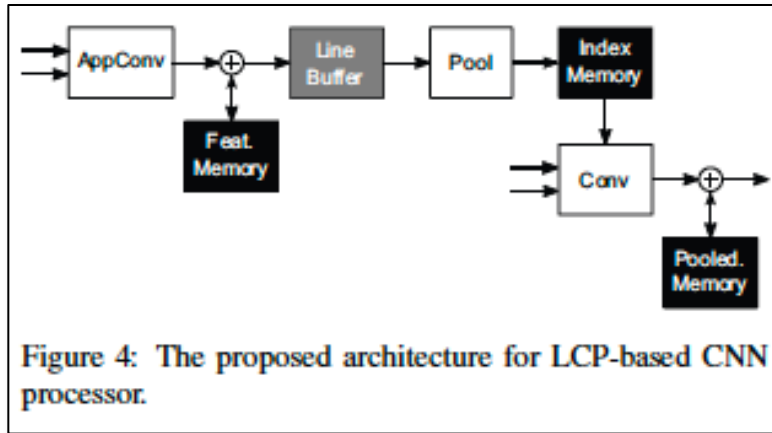
Id., Fig. 4. And for the reasons stated above, a POSITA would have been motivated to replace *Ujii*'s Sign Connect approximation method with *Moons*'s precision scaling computation.

92. This modification would have only required minor well known hardware and software architecture changes to implement another known approximation technique into *Ujiie*'s system to reap the benefits described above. A POSITA would have understood that modifications to *Ujiie*'s software approximation processes using Sign Connect would have been minimal to instead implement *Moons*'s method of adaptive precision using the MSBs of the input data and weights. *Ujiie* already teaches a preprocessing step to modify the representative values of the weights used in the approximation step (i.e., Sign Connect). This preprocessing step would have been easily modified to instead adjust the bit width of the input data and weight values prior to implementing the approximate convolution computation. And the result of this approximation would have similarly been used in the max pooling step to determine a maximum value.

93. Moreover, a POSITA would have recognized that *Ujiie*'s "multiplications of the sign of weights" would have been readily replaced by *Moons*'s minimum bit-width (i.e., MSB) computations at the AppConv step. As indicated by *Moons*, "[l]ower precision computing, i.e. computing using less bits...can be applied to a common digital multiplier." *Moons*, 5. *Ujiie* expressly provides support that "general multipliers" are typically used in the computation process. *Ujiie*, 4. Thus, it would have been well within the skill of a POSITA to incorporate well-known techniques, like *Moons*'s "[l]ower precision computing," in

Ujii's system to perform computations using the MSBs from the input data and weights. For example, the common multiplier described by *Ujii* could have instead been used to perform the decreased bit-width computations of *Moons*. It would have been well within the skill of a POSITA to modify the common multiplier to multiply the input data and weights using fewer bits rather than multiplying the input data by the weight value determined by Sign Connect. This would have simply required modifying the known method of multiplication via Sign Connect to instead perform the known method of multiplication using a decreased bit width. Indeed, *Ujii*'s system already performs computations using full bit width at its Conv step. Accordingly, *Ujii*'s system is capable of performing convolutional computations on values represented by more than one bit and therefore would have readily performed convolutions on a decreased number of bits as taught by *Moons*.

94. Additionally, *Ujii*'s hardware would allow for the use of MSBs in the approximate computation while retaining the full value in memory for use in the full precision computation once a max value has been determined. For example, *Ujii* teaches that the results of the AppConv and pooling are "temporarily stored in the index memory connected to the pooling circuit and utilized as the signal that controls inputs to the Conv circuit." *Ujii*, 4-5. This is shown in Figure 4 below:

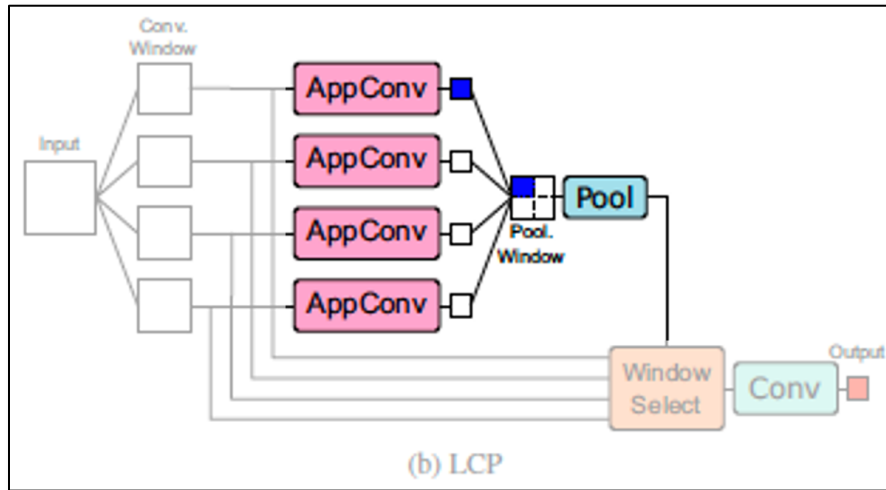


Id., Fig. 4. In the modified system, the approximate computations performed on the MSBs of the input data and weights would similarly be stored in the index memory. Then, full precision is performed on the data set that exhibited the maximum value from the approximate computations. This would have involved the same process as taught by *Ujiie*, whereby the result of the Max Pooling step would be “utilized as the signal that control inputs to the Conv” whereby, in the modified system, the resulting maximum value from *Moons’s* low bit width computation would be used to determine the data set to perform full precision convolution on at Conv.

5. Claim 1(b): “examining a first set of values computed for the plurality of data sets in the first iteration to determine whether a maximum value is present among the first set of values;”

95. *Ujiie-Moons* teaches computing a first set of values for the plurality of data sets in the first iteration. *Supra*, Claim 1(a). *Ujiie’s* system then examines the first set of values to determine whether a maximum value is present among the first

set of values in a “max pooling operation[.]” *Ujii*e, 1. *Ujii*e’s approximate max pooling operation is illustrated in Figure 2b below:



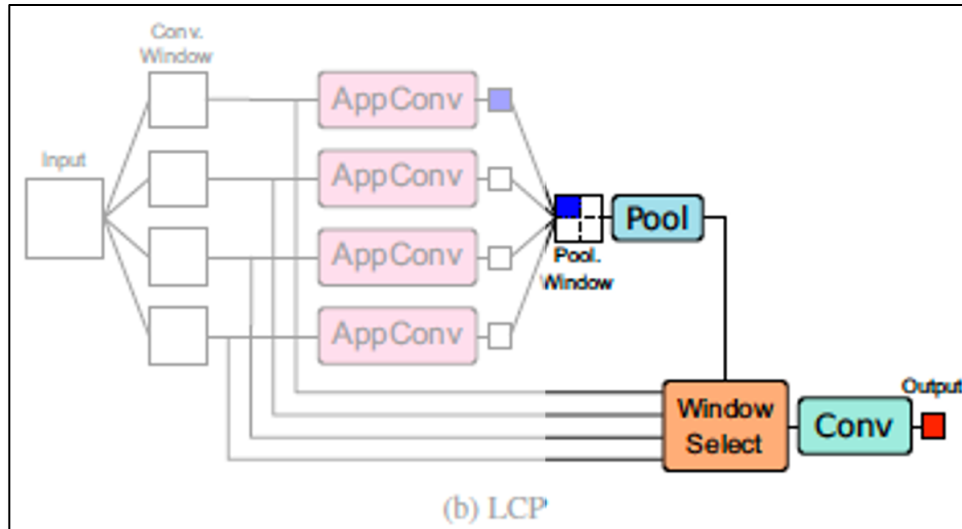
Id., Fig. 2b (modified for emphasis).

96. *Ujii*e teaches that it “use[s] max pooling, which selects the max value from the pooling window.” *Ujii*e, 2. In *Ujii*e’s LCP computation flow, “Step 1” computes “AppConv using the lightweight and approximate computation”—which, as modified by *Moons* calculates AppConv using precision scaling. *Id.*, 4. Next, *Ujii*e proceeds to “Step 2” to “operate max pooling to the results of AppConv to specify the convolution index of which the convolution result is propagated.” *Id.* In Figure 2b, *Ujii*e calculates four approximate AppConv results, examines those results using max pooling, and the largest value is determined as the maximum value, shown as a dark blue pixel. That is, *Ujii*e’s LCP flow specifies a particular window following AppConv. *Id.*, 3 (“LCP only executes an exact convolution (Conv) against **the specified window** predicted by the approximate convolution

(AppConv).”), 2 (“we use max pooling, which selects the max value from the pooling window.”). In this way, *Ujiie* “**adopt[s] the approximation only to predict the region to be selected in the max pooling**” (i.e., *determine[s] whether a maximum value is present among the first set of values*). *Ujiie*, 3.

6. **Claim 1(c): “responsive to identifying the maximum value, performing a full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value; and”**

97. *Ujiie* teaches the goal of its LCP computing method is to “predict[] a local region that will be **selected later in the pooling layer to limit the application of the accurate and thus expensive convolutional operations to the predicted region.**” *Ujiie*, 1. CNN convolution operations “are the most computationally intensive part in CNN.” *Id.*, 3; *see also, id.*, 1 (“...the accurate and thus expensive convolutional operations...”). Thus, *Ujiie* limits the expensive full precision computations to the dataset identified as a maximum from the max pooling step. *Ujiie*, 4. Indeed, in Step 3 of *Ujiie*’s LCP computing method, responsive to identifying the maximum, *Ujiie* “compute[s] Conv only to the input corresponding to the specified index in Step 2.” *Id.*, 4. This “Conv” computation is a “regular (and exact) convolution unit (Conv)[.]” *Id.* This exact (full precision) convolution is shown below in Figure 2b:



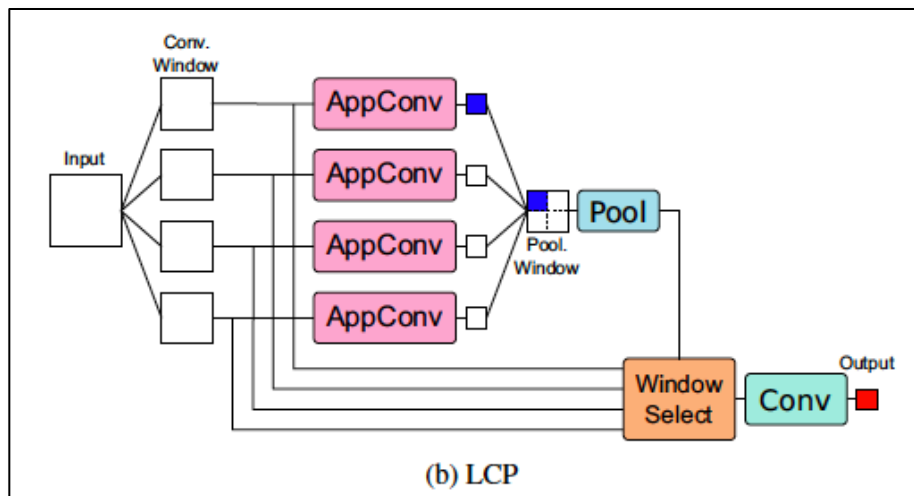
Id., Fig. 2(b) (modified for emphasis).

98. *Ujii*’s “regular” “exact” Conv convolution is a “full precision” convolution. First, “exact” convolutions require the full bit-width; anything less than the full bit-width would be an approximation. *Han* (EX1026), 8 (“In order to have absolute no loss of accuracy, switching to 32 bit arithmetic would not substantially affect the power or area of EIE.”); *Chippa* (EX1027), 3 (“Bit truncation is one such commonly employed technique [employing approximate data representations], where approximations are introduced by reducing the data width used to represent variables in hardware or software.”). *Ujii* discusses “conventional” convolutions, which a POSITA would have understood to be full precision, using 32-bit floating point representations (or the entire bit width of the binary value if not 32-bit floating point). *Judd*, Abstract (“A diverse set of CNNs is analyzed showing that compared to a **conventional implementation using a 32-bit floating-point** representation for

all layers,”). Second, regular, exact convolutions use all available bits (full precision) and are accurate, but computationally expensive. *Ujii*, 1. CNN convolution operations “are the most computationally intensive part in CNN.” *Id.*, 3; *see also, id.*, 1 (“...the accurate and thus expensive convolutional operations...”). That is why *Ujii*’s goal is to limit the number of exact, full precision convolutions to the most important datasets exhibiting the max from the Step 2 max pooling operation. *Id.*, 1 (“LCP predicts a local region that will be selected later in the pooling layer **to limit the application of the accurate and thus expensive convolutional operations to the predicted region.**”).

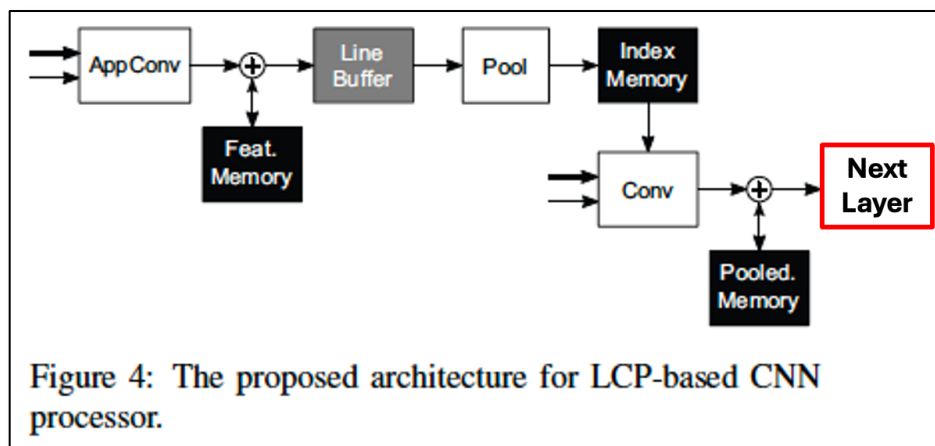
7. Claim 1(d): “propagating the full precision computation of the value to a subsequent layer of the CNN.”

99. Once *Ujii*’s system computes a full precision computation on the dataset that exhibited the maximum, that full precision value is propagated as output to the next layer. *Ujii*, 3. As a first example illustrating *Ujii*’s propagation of the full precision value, Figure 2b illustrates the “output” of the Conv step as a red pixel:



Id., Fig. 2b. The red pixel is the full precision “output” and is passed to subsequent layers of the CNN processor.

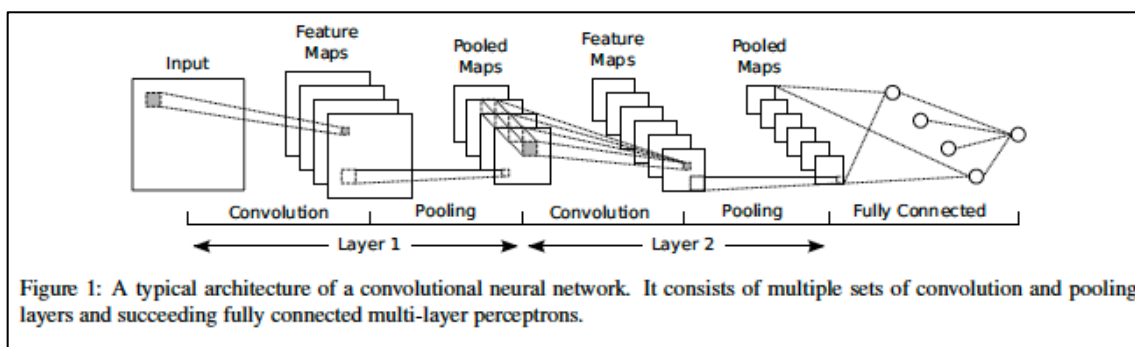
100. As a second example illustrating *Ujii*e’s propagation of the full precision value, *Ujii*e provides an overview of its proposed architecture in Figure 4:



Id., Fig. 4 (annotated). The final arrow, annotated in Fig. 4 (above), shows that the Conv results stored in pooled memory are sent to a subsequent layer. As discussed in further detail below, this would inform a POSITA that *Ujii*e is propagating the

convolution result to a next layer, including another convolution and pooling layer, and eventually to a fully connected layer.

101. *Ujii* exemplifies this process in Figure 1, showing how in typical CNNs, the output of convolution/pooling is propagated to “Layer 2” of the CNN for additional convolution and pooling steps or to a “Fully Connected” layer.



Ujii, Figure 1; *see also id.*, 3 (“Figure 1: A typical architecture of a convolutional neural network. It consists of multiple sets of convolution and pooling layers and succeeding fully connected multi-layer perceptrons”).

102. As a third example illustrating *Ujii*’s propagation of the full precision value, *Ujii*’s Tables 1 and 2 indicate the network parameters used to evaluate the classification accuracy of the LCP-based CNN processor. Table 1 indicates that additional layers, including ReLU, fully connected, and softmax, are used in analyzing the MNIST dataset. Table 2 indicates that additional layers, including ReLU, convolution, fully connected, and softmax, are used in analyzing the CIFAR-

10 dataset. Each of these additional layers utilize the propagated values from the max pooling layer as indicated in Tables 1 and 2 below:

Layer name	Parameter	
Input image	size: 28x28, channel: 1	
Convolution	kernel: 5x5, channel: 20	
Max pooling	kernel: 2x2, stride: 2	
ReLU		
Convolution	kernel: 5x5, channel: 50	
Max pooling	kernel: 2x2, stride: 2	
ReLU		Next Layers
Fully connected	channel: 500	
ReLU		
Fully connected	channel: 10	
Softmax		

Layer name	Parameter	
Input image	size: 32x32, channel: 3	
Convolution	kernel: 5x5, channel: 64, padding: 2	
Max pooling	kernel: 2x2, stride: 2	
ReLU		
Convolution	kernel: 5x5, channel: 128, padding: 2	
Max pooling	kernel: 2x2, stride: 2	
ReLU		Next Layers
Convolution	kernel: 5x5, channel: 256, padding: 2	
Max pooling	kernel: 2x2, stride: 2	
ReLU		
Convolution	kernel: 5x5, channel: 512, padding: 2	
Max pooling	kernel: 2x2, stride: 2	
ReLU		
Convolution	kernel: 5x5, channel: 512, padding: 2	
Max pooling	kernel: 2x2, stride: 2	
ReLU		
Fully connected	channel: 512	
ReLU		
Fully connected	channel: 10	
Softmax		

Ujiie, Table 1 (top, annotated) and Table 2 (bottom, annotated). Therefore, *Ujiie*'s system propagates the full precision computation of the value (i.e., exact convolution output) to a subsequent layer of the CNN—e.g., “Layer 2[,],” including convolution and pooling, or the Fully Connected layer of *Ujiie*'s CNN.

8. Claim 10: The one or more non-transitory computer-readable storage media of claim 1, wherein the CNN is employed to analyze an image.

103. *Ujiie* teaches the one or more non-transitory computer-readable storage media of claim 1. *Supra*, Claim 1.

104. *Ujiie* teaches “[t]his paper proposes an efficient computing method, LazyConvPool (LCP), and its hardware architecture to reduce power consumption of **CNN-based image recognition**.” *Ujiie*, Abstract. *Ujiie* proposes an “experimental evaluation using **image classification dataset[s]**” such that its approximated prediction strategy for CNNs reduces power consumption for analyzing image data. *Id.*; *see also id.*, 5 (“We evaluated the classification accuracy of LCP with image recognition datasets. We used two datasets: the handwritten digits dataset MNIST [] and the object recognition dataset CIFAR-10”); Table 1 (“input image”), Table 2 (“input image”). *Ujiie* teaches that through “the experimental evaluation, the proposed LazyConvPool algorithm and its hardware architecture successfully reduce the energy consumption for **image recognition** by 11.4%–14.1% while keeping the recognition accuracy.” *Ujiie*, 6.

9. Claim 11: The one or more non-transitory computer-readable storage media of claim 1, wherein:

105. *Ujiie* teaches the one or more non-transitory computer-readable storage media of claim 1. *Supra*, Claim 1.

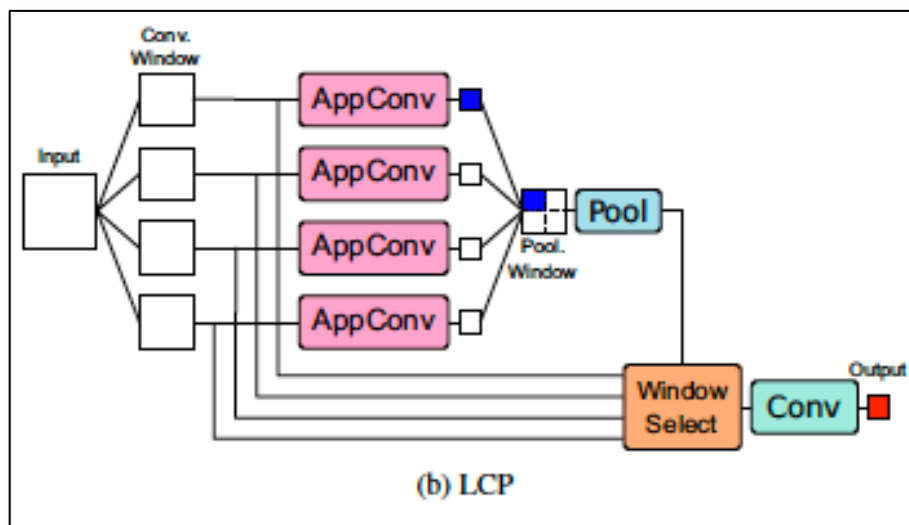
- a. **Claim 11(a): “the first iteration computes a value that approximates the full precision computation of the value; and”**

106. *Ujiie* teaches computing a value via AppConv that approximates the full precision computation of the value. *Supra*, Claim 1(a); *see also Ujiie*, 4 (“we construct LCP framework with two main parts: **(1) approximated convolution units (AppConv) to predict the feature selection** in the pooling layer, and **(2) a single regular (and exact) convolution unit (Conv) to perform convolution only on the predicted window.**”). As modified by *Moons*, this approximation involves quantizing the input data and weights. *Supra*, Claim 1(a).

- b. **Claim 11(b): “the full precision computation is performed on the data set that includes less data than the plurality of data sets.”**

107. *Ujiie*’s system only performs full precision computations on the data set whose approximated value exhibits the maximum. *Supra*, Claims 1(b)-(c); *Ujiie*, 3 (“**LCP only executes an exact convolution (Conv) against the specified window predicted by the approximate convolution (AppConv).**”). As shown in Fig. 2b of *Ujiie*, four convolution windows are analyzed by a respective AppConv computation. Each of the four outputs of AppConv are passed to the pooling window, where a single convolution index associated with the maximum value is selected. *Ujiie*, 4 (discussing Step 1 and Step 2). The full Conv computation is performed on the “input corresponding to the specified index in Step 2.” *Id.*,

(discussing Step 3). By limiting the full precision computation to the data set exhibiting the maximum approximated value, the full precision computation is performed on less data than the plurality of data sets. Said another way, the full precision computation is performed on the data set that includes less data than the plurality of data sets because, in its example, *Ujiie* does not perform full precision computation on three of the four data sets (indicated with white cells in Fig. 2b) but rather only on the maximum approximated value. This maximum approximated value is indicated by the dark blue pixel in Figure 2b:



Ujiie, Fig. 2b.

VII. OPINION: GROUNDS 2 AND 3

A. Obviousness over Ujiie, Moons, and Kaul, or Alternatively Obvious over Ujiie, Moon, Kaul and Ishii

1. Overview of Kaul

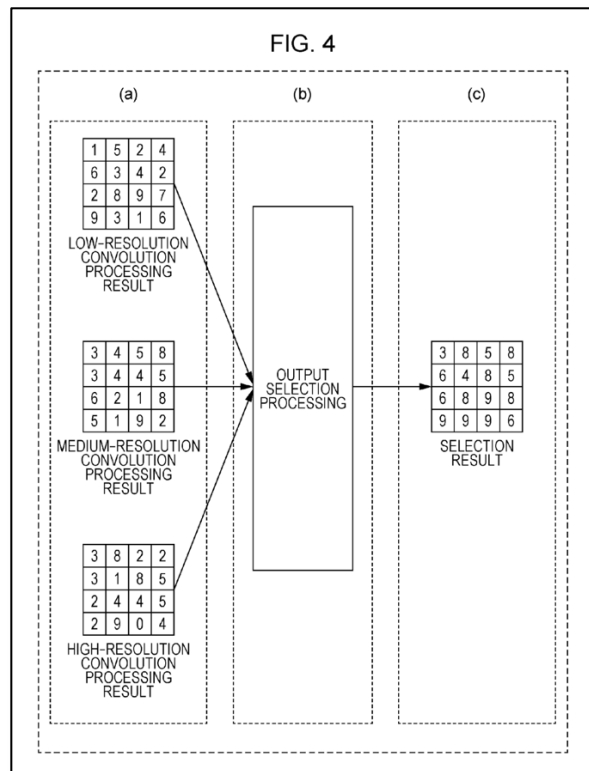
108. *Kaul* is generally directed to “[a]daptive precision” for “[e]nergy-efficient k-nearest-neighbor (kNN) computations [which] are key building blocks for computer vision, classification, and machine-learning workloads[.]” *Kaul* (EX1006), 1. Using adaptive precision, *Kaul* “improves energy efficiency by eliminating a majority of vectors without costly full-precision computation, with as-needed precision refinement[.]” *Id.* To this aim, *Kaul*’s system uses “MSB-to-LSB-based selective distance refinement” which “start[s] with low-accuracy from the MSB” and then performs “[i]terative refinement of accumulated distance...as needed with successive sort operations on 2b windows from MSB to LSB.” *Id.* This method allows for “[e]arly removal of vectors while moving from MSB to LSB for nearest-neighbor search result[ing] in 47% compute reduction[.]” *Id.*

2. Overview of Ishii

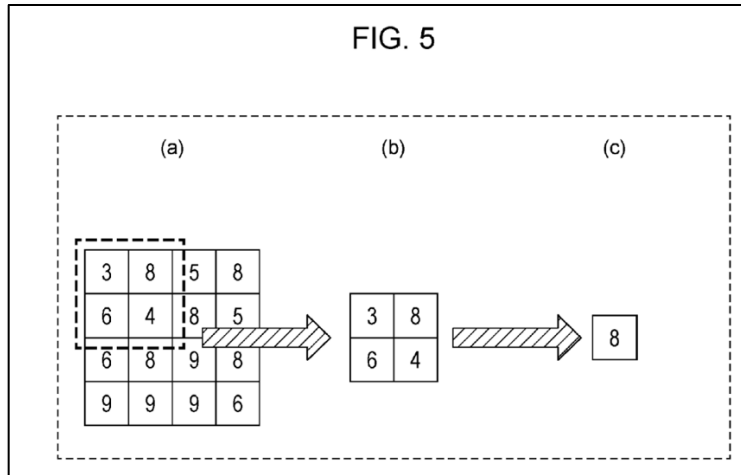
109. U.S. Patent Application Publication No. 2016/0259995 to Ishii et al. (“*Ishii*”) was filed on February 22, 2016, and published on September 8, 2016.

110. *Ishii* is generally directed to “an image recognition method and the like that enables high-speed and high-precision image recognition.” *Ishii*, [0006]. *Ishii*’s image recognition system includes “convolution processing” and “subsampling [i.e.,

“Max-pooling”] processing.” *Id.*, [0059], [0073]. *Ishii*’s convolutional processing includes “low-resolution[,]” “medium-resolution[,]” and “high-resolution” processing, where the system “select[s] a processing result having a maximum value from among a plurality of values” from the various convolutions. *Id.*, [0102], [0119], Fig. 4.



Id., Fig. 2. “[I]n a case where there are two or more maximum values, the output selecting unit 124 need just select any one of the two or more maximum values.” *Id.*, [0130]. *Ishii* similarly teaches that in the event of max-pooling, if two or more values are the same, the system will proceed with one of those values. *Id.*, [0142], Fig. 5.



Id., Fig. 5.

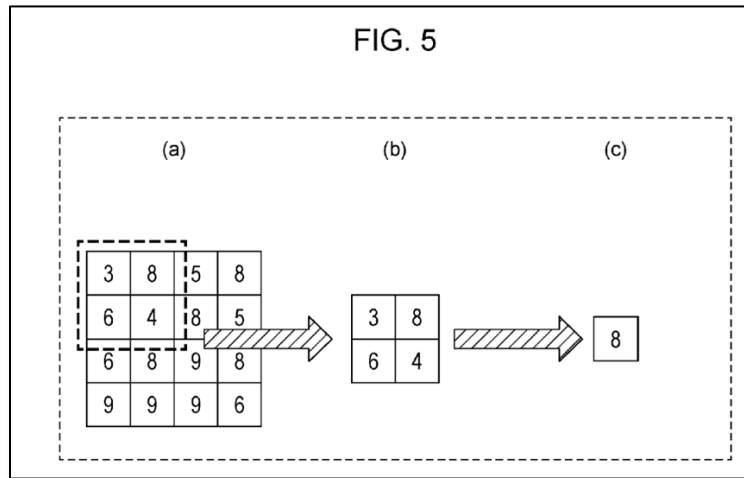
3. **Claim 2: The one or more non-transitory computer-readable storage media of claim 1, further comprising: responsive to determining that the first set of values are the same, performing, by the at least one processor, a second iteration that includes computing the value based on a second set of MSBs for each of the plurality of data sets, the second set of MSBs being larger than the first set of MSBs.**

111. *Ujiie* teaches the one or more non-transitory computer-readable storage media of claim 1. *Supra*, Claim 1.

a. Ujiie-Moons in View of Kaul

112. *Ujiie-Moons* teaches performing an iteration that computes an approximated value based on MSBs and then uses max pooling to identify a single dataset that exhibits the maximum to perform a full precision convolution on. *Supra*, Claim 1(a)-(c). However, *Ujiie-Moons* does not discuss what happens when max pooling does not identify a maximum. For example, when the output approximated results for two or more datasets are the same. *See, e.g., Ishii* (EX1007), [0130],

(“Note that in a case where there are two or more maximum values, the output selecting unit 124 need just select any one of the two or more maximum values of the processing results”), [0142] (describing selecting maximum after max pooling despite a tie), Fig. 5 (depicting the same).



Id., Fig. 5. Similar tie-breaking strategies were well-known in the prior art. For example, *Watanachaturaporn*’s (EX1028) tie-breaking strategy was to “randomly select one of the class labels that are tied.” *Watanachaturaporn*, 3 (“The output from each classifier is obtained in the form of a class label. The class label that occurs the most is assigned to that pixel. A tie-breaking strategy is adopted in case of a tie. A common tie-breaking strategy is to randomly select one of the class labels that are tied.”). *Wen* studied neuromorphic computing systems using memristors with crossbar topologies. In the event of ties in determining routing order, *Wen* looked to wire weighting as the tie breaker. *Wen* (EX1029), 5 (“The routing order is determined by the distance from the center of gravity of all cells to its closest pin of

wires. If the distance is the same for more than two wires, we will use wire weighting as the tie breaker.”). Finally, *Tahir* discloses two tie-breaking strategies: 1) arbitrarily assigning a tied pixel or 2) assigning the class with another attribute (e.g., competitive distance) exhibiting a maximum value. *Tahir* (EX1032), 5 (“Case 4 (Tie-Break): If the competitive distances of any two classes have same value and this value is greater than the competitive distance of the other class, [...] then assign one of the classes arbitrarily to the pixel of the final output image, otherwise assign the class with maximum competitive distance to the pixel of the final output image.”).

113. A POSITA would have understood in *Ujiie-Moons*, if two datasets share the same approximated value in a max pooling process, a single dataset cannot be accurately selected for a full precision convolution under *Ujiie*’s native prediction strategy. Specifically, *Ujiie* is concerned with specifying the convolution “index” that points to the underlying convolution that will be computed and propagated, thus computing the convolution only for the “specified index[.]” *Ujiie*, 4 (discussing Step 2 and Step 3). A POSITA would have looked to known techniques to resolve shared approximated values following max pooling to identify a maximum enabling *Ujiie*’s prediction strategy to select a single dataset representing the most significant input data for full precision convolution. That is, selecting an index would have required the selection of a specific value associated with a specific index. In the instance of

two or more shared values, *Ujii*'s index selection would have necessitated techniques to specifically identify an index. *Kaul* provides a known technique to solve this issue. *Kaul* teaches iteratively increasing bit width in neural network computations when a first iteration does not provide an answer.

114. *Kaul*, similar to *Ujii* and *Moons*, teaches an “[a]daptive precision” method for computer vision that “improves energy efficiency by eliminating a majority of vectors without costly full-precision computation, with as-needed precision refinement to guarantee...accuracy[.]” *Kaul*, 1. Specifically, *Kaul*'s method “start[s] with low-accuracy from the MSB using partial distance compute circuits...[then] refine[s] the remaining distances, with a successive 2b [2 bits] of the accumulated distance available for further sorting in each iteration.” *Id.* This “[i]terative refinement of accumulated distance is interleaved as needed with successive sort operations on 2b windows from MSB to LSB.” *Id.* *Kaul*'s iterative process is illustrated below:

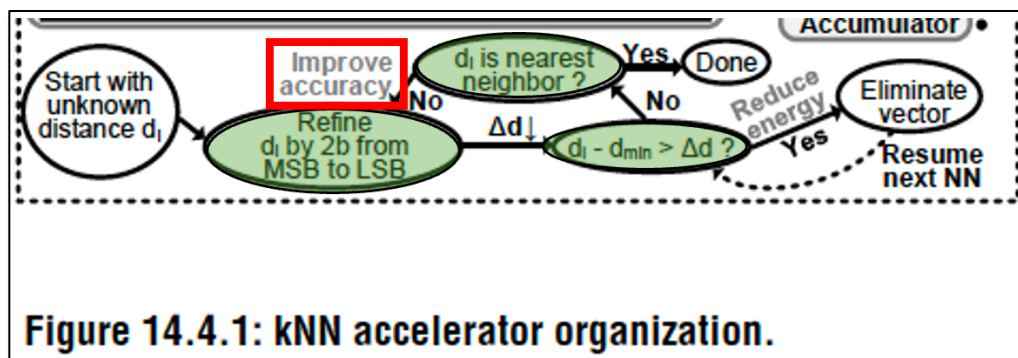
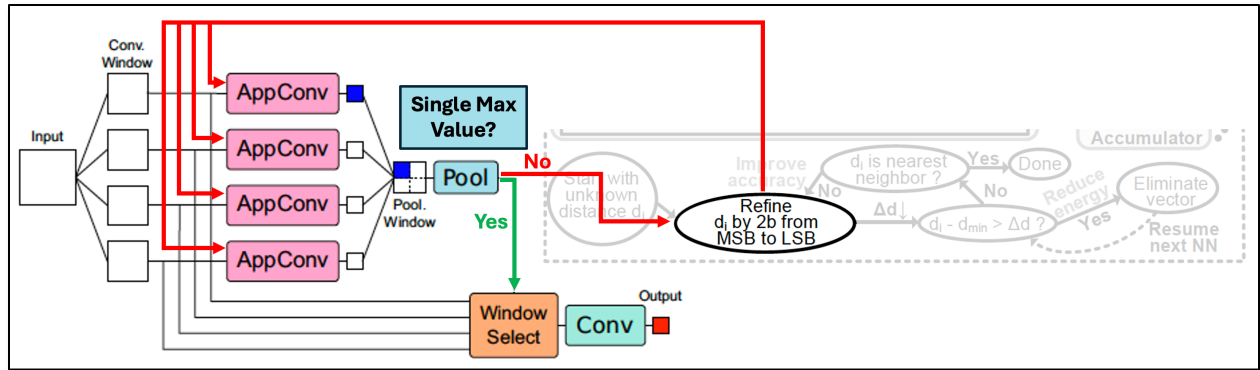


Figure 14.4.1: kNN accelerator organization.

Id., Fig. 14.4.1 (excerpted and annotated).

115. As shown in Fig. 14.4.1, *Kaul's* method determines if the nearest neighbor was found using the current bit-precision, and if not, “[i]mprove[s] accuracy” by refining “by 2b [2 bits] from MSB to LSB[.]” *Id.*, Figure 14.4.1. This is further highlighted in the annotated Fig. 14.4.1 above. After refining d_i by 2b from MSB to LSB (the bottom-left green oval), a check is performed on whether $d_i - d_{min} > \Delta d$. If not, then a check is performed on whether d_i is the nearest neighbor. If the nearest neighbor is not found, *Kaul* performs a second iteration of computations that includes two additional bits.

116. In *Ujii-Moons-Kaul*, *Ujii's* modified system performs a first iteration of approximate convolutions using a first number of MSBs and performs max pooling. As discussed above at ¶113, *Ujii* determines whether a singular max value exists to determine the singular index to propagate to the subsequent Conv step. If no maximum is found (i.e., *responsive to determining that the first set of values are the same...*), then *Ujii-Moons-Kaul* performs a second iteration pursuant to *Kaul* with two additional bits (i.e., “performing a second iteration that includes computing the value based on a second set of MSBs for each of the plurality of data sets, the second set of MSBs being larger than the first set of MSBs”). Below is an illustration of the *Ujii-Moons-Kaul* modified system.



Ujii, Fig. 2b; *Kaul*, Fig. 14.4.1 (annotated and modified for emphasis).

i. Motivation to Combine Ujii-Moons and Kaul

117. A POSITA would have understood that by using decreased approximated precision before max pooling, as taught by *Ujii-Moons*, a clear maximum may not always be determined. This would occur, for example, when analyzing an area of the image with minimal differences with the surrounding area(s). Such areas are commonplace within images. This includes facial recognition, where there are many images taken whereby the background behind the individual is fairly “flat” with minimal variations. As another example, medical imaging of a patient’s skin may have a very similar tone throughout the image besides the area of interest, which may be of a different color. In these exemplary scenarios, the pixel values representing the background regions of a similar color have very similar pixel values. Therefore, without a large difference in pixel values, when approximating the convolution, the resulting values of these regions for max pooling would potentially result in the same value.

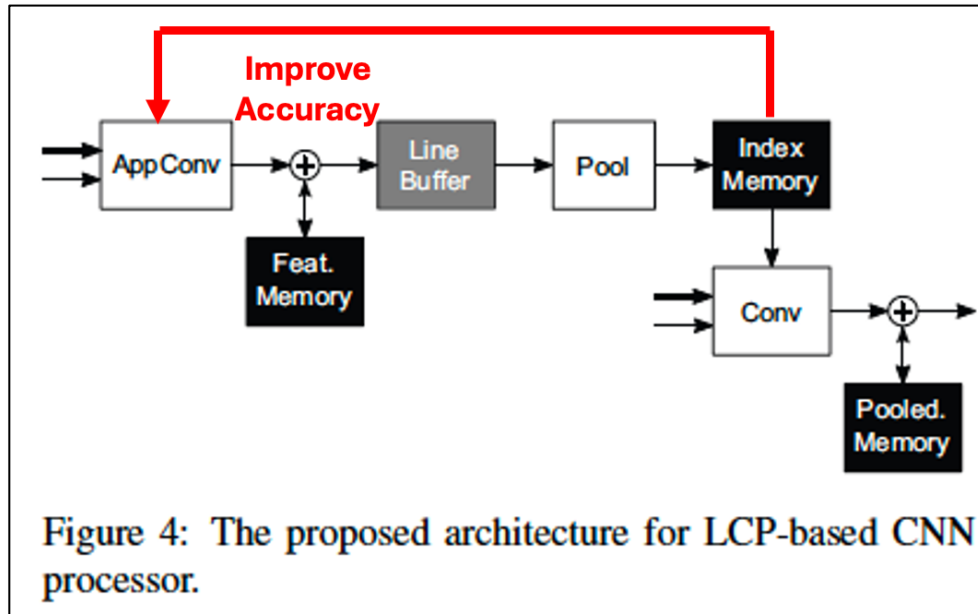
118. A POSITA would have understood that where the AppConv process identifies no maximum value—such as a tie described in *Ishii*—it would have been obvious in view of *Kaul* to iterate by computing a higher precision approximation to distinguish between the similar pixel input data to identify which input data, if any, produces a maximum value. For example, a POSITA would have understood that using such approximation would result in scenarios in which regions having similar pixel values would have convolutional results of the same values. As such, for the system to accurately determine the most significant input data, the approximation step, if performed at greater precision (i.e., with more bits), would identify any data that produces a maximum value. As such, the system would benefit by properly iterating the approximate convolution with greater precision to proceed with the data set representing the most significant input data for full precision convolution.

119. This would have increased the accuracy of the AppConv process without substantially increasing energy consumption. *Kaul*, 1 (“Adaptive precision improves energy efficiency...with as-needed precision refinement to guarantee kNN accuracy of closely matched vectors”), *see also id.* (describing shorter bit-width also “reduces...latency[,]” decreases storage area usage, and “reduces [energy usage] up to 3x compared to a full-precision design”). For example, iterating with two more bits would support a more precise output value while striking a balance for energy expenditure, latency, and storage. The process of performing the first iteration with

the minimal amount of bits that is typically required to determine a maximum would decrease energy consumption to its lowest necessary levels. And in the scenario where no maximum value is found, performing the second iteration with more, but not all, bits would increase the likelihood of identifying a maximum while still decreasing energy consumption as compared to a full precision computation. By including more bits in each of the input data and weight values, the likelihood of identifying a maximum value on the second iteration would have increased significantly. Therefore, it would have been obvious to perform the second iteration with a minimal amount of increase in bit width because jumping to a full precision computation on the second iteration would unnecessarily increase energy consumption when a lower precision computation would likely have identified a maximum.

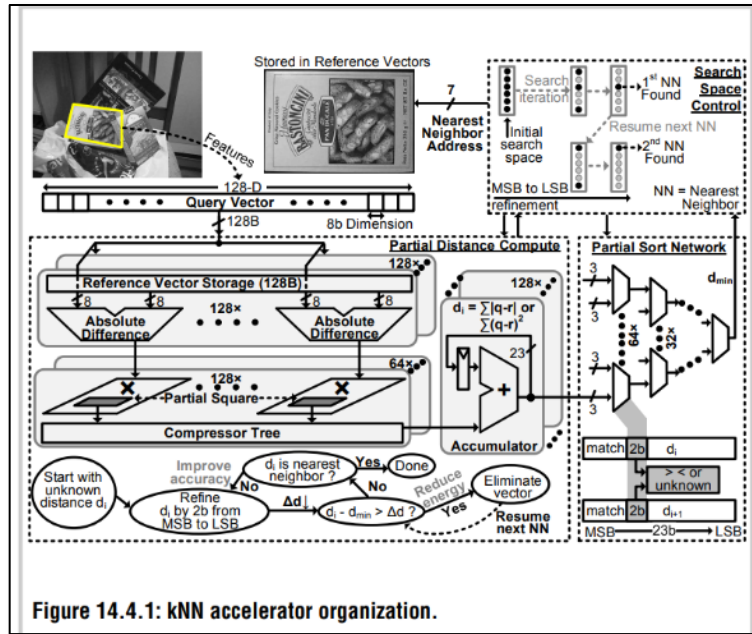
120. Moreover, this would have ensured *Ujii*'s native prediction strategy would find a single max dataset, performing iterations until a delta between datasets is met, exhibiting a maximum. In other words, this iterative process would almost always reach a conclusion of a maximum value because, eventually a full precision convolution (i.e., all bits per value) would be performed, assuming no maximum value had been determined during the less precise convolutions.

121. A POSITA would have appreciated a reasonable expectation of success modifying *Ujiie-Moons* to perform *Kaul's* adaptive precision technique. *Ujiie's* annotated Figure 4 below shows an example of how this process would work:



Ujiie, Fig. 4 (annotated).

122. As shown, *Ujiie-Moons's* system performs a normal iteration of low bit precision AppConv and max pooling. And in the modified system, this would include performing the approximate convolution computations for each of the data sets with minimal bits. The results of the pooling operation are accumulated in Index Memory, and the results are analyzed for the maximum to determine which dataset receives full-precision Conv. If no maximum value exists (e.g., two datasets tied) on the first iteration, the process proceeds back to AppConv for a second iteration using an increased number of MSBs, as taught by *Kaul*. *Kaul*, 1, Fig. 14.4.1.



Id., Fig. 14.4.1. Specifically, the system would determine that two or more values in the max pooling step exhibit the maximum value found among the four values. As such, no maximum value exists. Therefore, the system would iterate the approximate convolution, including more bits as taught by *Kaul*.

123. The system would then perform the second iteration of convolution and max pooling with the weights and input containing more MSBs than the first iteration. As described above, this would allow for a greater precision convolutional computation which would significantly increase the likelihood of identifying a maximum value. *Ujii*'s memory would allow for retrieval of the original data in its full precision format to allow for modification of the bit width as taught by *Moons*. Namely, *Ujii*'s system stores the original data in its full precision format readily available to select for the Conv step. As such, the system would readily pull from

memory to retrieve the full precision weights and input data and decrease the bit width of each to perform a second iteration of approximate convolutions.

124. A POSITA would have understood that *Kaul's* kNN-based adaptive precision technique is compatible with *Ujiie's* CNN processing. For example, convolution computations in CNNs and kNN vector comparisons both involve multiplication and addition across vectors. For CNNs, the process involves transforming images into feature vectors through successive convolutional and pooling operations. See § V.A. Linear algebraic operations are used in each of these layers, which includes multiplication and summation of binary values. Similarly, for kNN analysis, the image is represented as a feature vector and, when classifying the image, its feature vector is compared to feature vectors of labeled examples. These operations include distance metrics (e.g., Euclidean distance), which similarly involve computations of binary values in vector form.

125. In this case, the precision of the values within the vector (i.e., each value's bit width) would affect the computations in essentially the same way. The algebraic vector-based computations would similarly have increased precision with increased bit width of the values and, as a tradeoff, expend more computational energy performing the higher precision computations. As such, the same bit width implications for computational efficiency would apply to both. Specifically for the present analysis, both would operate at lower bit width values in the vectors for the

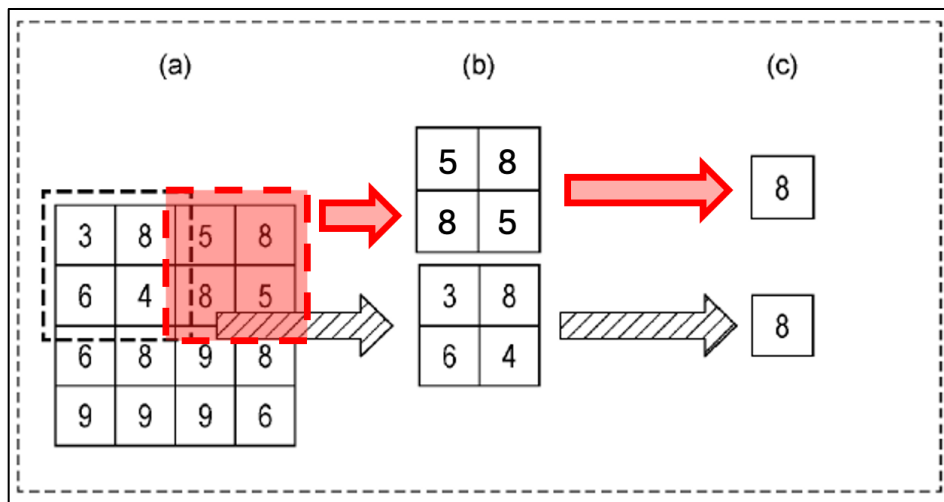
initial computations and then, when a resolution is undeterminable (e.g., no maximum value in CNNs or no nearest neighbor in kNNs), increase precision by including more bits in the vector values. As discussed above, this iterative process would increase the likelihood of determining a maximum value in CNN max pooling as well as determining the nearest neighbor in kNNs. Therefore, a POSITA would have realized that adaptive precision would be readily applied to vector-based computer vision computations, including in both CNNs and kNNs.

b. Ujiie-Moons-Kaul in View of Ishii

126. *Ujiie-Moons-Kaul* teaches a process to identify the single dataset that receives a full precision convolution when two or more datasets share the same approximated convolution value by iteratively increasing bit-widths for the approximated computations. *Supra*, §VII.A.3.a. *Ujiie's* native prediction strategy requires identifying a maximum. Specifically, *Ujiie* is concerned with specifying the convolution “index” of which the convolution result is propagated, then computing the convolution only to the “specified index[.]” *Ujiie*, 4 (discussing Step 2 and Step 3). A POSITA would have understood that the “index” refers to a specific location within the dataset. Therefore, *Ujiie's* process necessarily determines if the approximated values exhibit a maximum, or are the same and thus fails to identify a maximum. *Ujiie's* modified system therefore necessarily performs a computational check whether the values are the same. For example, *Ujiie* selects an index of a single

dataset to propagate for full convolution. *Ujiie*, 4 (discussing Step 2 and Step 3). Selecting an index would have required the selection of a specific value associated with a specific index. This computational check was well-known in the art, as evidenced by *Ishii* above. For example, *Ishii* discloses an “output selecting unit” to “select[] a processing result having a maximum value,” then “determin[ing] the selected value of the processing result[.]” *Ishii*, [0128]. Thus, it would have been obvious to incorporate *Ishii*’s express teachings into *Ujiie*’s modified system, as I discuss in more detail below.

127. *Ishii* teaches a “classifier 90a includes an image input unit 91 that reads an image (input image), a convolution processing unit 92 that performs convolution processing on the input image, [and] a subsampling unit 93 that performs subsampling [i.e., max-pooling] processing[.]” *Ishii*, [0059], [0073] (“selecting a maximum value by subsampling is called Max-pooling”). *Ishii* demonstrates a scenario in which the results of max-pooling include a tie between two values. For example, *Ishii* teaches the subsampling processing includes sampling a 2x2 region “for selecting a maximum pixel value[.]” *Id.*, [0142]. *Ishii*’s examples include both a scenario in which the maximum is straightforward (i.e., between “pixel values 3, 8, 6, and 4”) and scenarios in which there is a tie (e.g., in the “2x2 region including an upper right pixel” as shown in Figure 5, showing values 5, 8, 8, and 5). *Id.* Figure 5 is reproduced below:

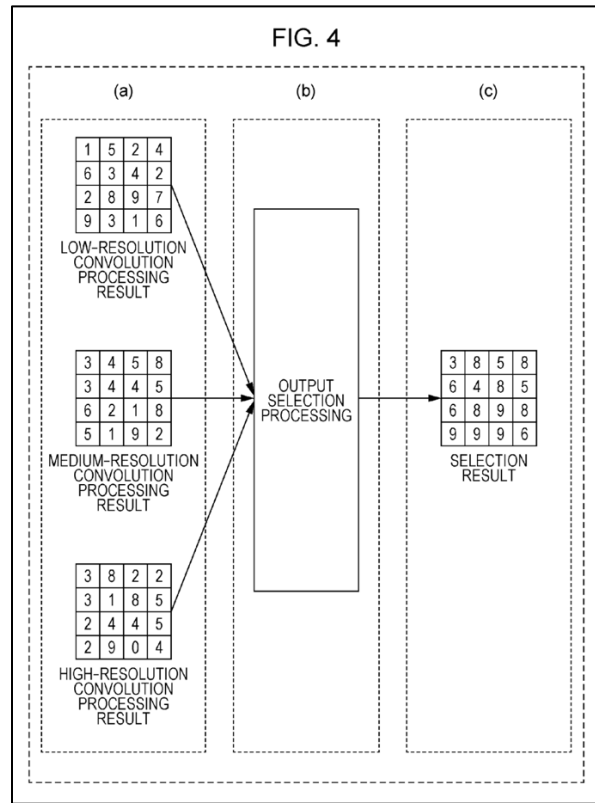


Id., Fig. 5 (annotated).

128. As shown in the annotated Fig. 5 above, two or more values exhibit the maximum (e.g., the value “8”). *Ishii* teaches or at least suggests that its system would choose between one of the two values exhibiting the maximum, such that the system “selects 8...as representative...[of the] 2x2 region including [the] upper right pixel[.]” *Ishii*, [0142]. A POSITA would have understood that *Ishii*’s disclosure of selecting the value “8” (see Fig. 5c, above) from the values 8, 8, 5, and 5 (see Fig. 5b, above), that *Ishii* is necessarily choosing between one of the two instances of the value “8.” *Ishii* thus teaches or suggests a computational check to identify a maximum, such that if there is a tie, *Ishii*’s system selects one to proceed to the next layer.

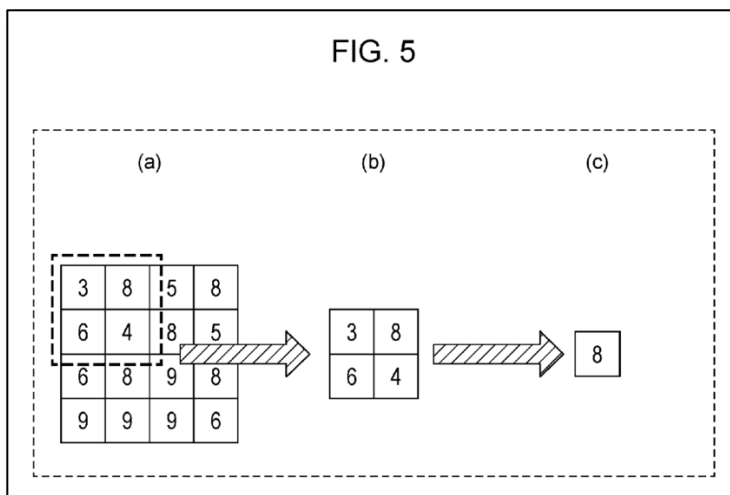
129. Indeed, *Ishii* teaches a similar selection process for its convolutional computations, where *Ishii* “selects a maximum value from among values...of three processing results of the convolution processing” including results from “low-

resolution convolution[.]” “medium-resolution convolution[.]” and “high-resolution convolution[.]” *Ishii*, [0119], Fig. 4.



Id., Fig. 4. In determining its maximum here, *Ishii* teaches “in a case where there are two or more maximum values, the output selecting unit 124 need just select any one of the two or more maximum values of the processing results.” *Id.*, [0130]. A POSITA would have understood or at least found obvious that this same tie-break process for same-value convolutions would have been used to tie-break *Ishii*’s max pooling process (i.e., the processor “select[s] any one of the two or more maximum values”). As discussed immediately above, for the top-right 2x2 region with values of 8, 8, 5, and 5, *Ishii* expressly discloses that 8 is chosen as the representative

feature. Thus, a POSITA would have understood that *Ishii*'s disclosure regarding the tie-break process discussed in [0119] regarding Fig. 4 would also apply to a similar tie-break scenario discussed in [0142] regarding Fig. 5. In the case of Fig. 5, the sub-sampling unit 13, "need just select any one of the two or more maximum values" from the selection results of Fig. 4. *Ishii*, [0119], [0130], [0142].



Ishii, Fig. 5.

i. Motivation to Combine Ujiie-Moons-Kaul and Ishii

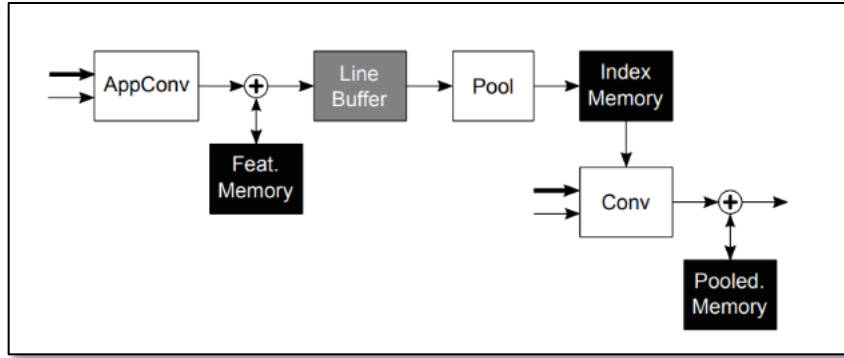
130. A POSITA would have been motivated and found it obvious to modify *Ujiie-Moons-Kaul* to perform computational checks on the maximum value of *Ujiie*'s max pooling operation to determine if two or more values exhibited the maximum, as taught by *Ishii*. Namely, a POSITA would have understood the low precision approximate computations in *Ujiie*'s modified system would cause scenarios where two or more values exhibit the maximum. As described above, decreasing the precision of the convolution in areas of an image that are mostly

uniform (e.g., same or similar background color) will result in approximated input data values that are similar and therefore more likely to provide two or more maximum values. *Supra*, §VII.A.3.a. As such, a POSITA would have been motivated to implement *Ishii*'s computational check in *Ujiie*'s modified system to identify this scenario and resolve it. For instance, it would have been obvious that selecting randomly from two or more values exhibiting a maximum would have, at least in some instances, caused the system to select the incorrect data set to perform full precision computation on. Therefore, a POSITA would have been motivated to prevent this error from propagating through the next CNN layers thereby increasing the likelihood of misclassification. Indeed, a POSITA would have understood, as did *Ujiie*, that a goal of *Ujiie*'s CNN is to prevent "approximation errors [that] are accumulated and inflated during the propagation." *Ujiie*, 3. And a POSITA would have recognized that identifying a scenario of two or more maximum values, as taught by *Ishii*, and then iterating to enhance precision, as taught by *Kaul*, would have prevented this.

131. To be clear, the proposed combination does not simply choose one of the tie value datasets as in *Ishii*. Instead, as set forth below, the combination adds bits and re-computes the AppConv values to identify which of the original datasets produces a maximum value.

132. A POSITA would have understood *Ujiie*'s native prediction strategy that mandates a single dataset to exhibit the max for full precision convolutions to teach, suggest, or motivate the combination with a reference that teaches a resolution if a maximum is not found, such as in *Ishii*. Using *Ishii*'s random selection tie-break functionality is, however, error prone. For instance, by randomly selecting one of two choices, there's a fifty percent chance of error. This is because in *Ujiie*'s modified system, the two approximated maximum values would likely have refined values that would identify a maximum between the two if the system increased the precision of the convolution. Therefore, selecting at random between the two would, in some instances, identify a data set not representative of the most significant input data to perform full precision on (i.e., roughly 50% of the time if two values exhibit the maximum). A POSITA would have instead used *Kaul*'s technique to increase bit-widths rather than *Ishii*'s random selection to increase system accuracy and for the reasons above.

133. A POSITA would have appreciated a reasonable expectation of success modifying *Ujiie-Moons-Kaul* to check whether two or more values exhibit the maximum. *Ujiie*'s modified system already processes approximate convolution results and stores them in its "Index Memory[.]" *Ujiie*, 4-5, Fig. 4.



Id., Fig. 4. Thus, this modification would have only required minor software changes in *Ujiie*'s system to check for a tie between these values for the benefits described above. For example, modifying the system to identify when two or more values exhibit the maximum value at *Ujiie*'s max pooling stage would have required simple programming that utilizes the comparison function already present in *Ujiie* to prevent forward propagation of a value when two or more values exhibit the maximum. Instead, the simple programming changes would direct the system to revert back to the data sets used for the approximation and perform the same computation using more bits than in the first iteration. Such programming modifications were well within the skill of a POSITA at the time of the '831 Patent.

4. Claim 3

- a. **Claim 3(a): "The one or more non-transitory computer-readable storage media of claim 2, further storing instructions, which, when executed by the at least one processor, cause the at least one processor to perform operations comprising: examining a second set of values computed for the plurality of data sets in the second iteration to determine whether the**

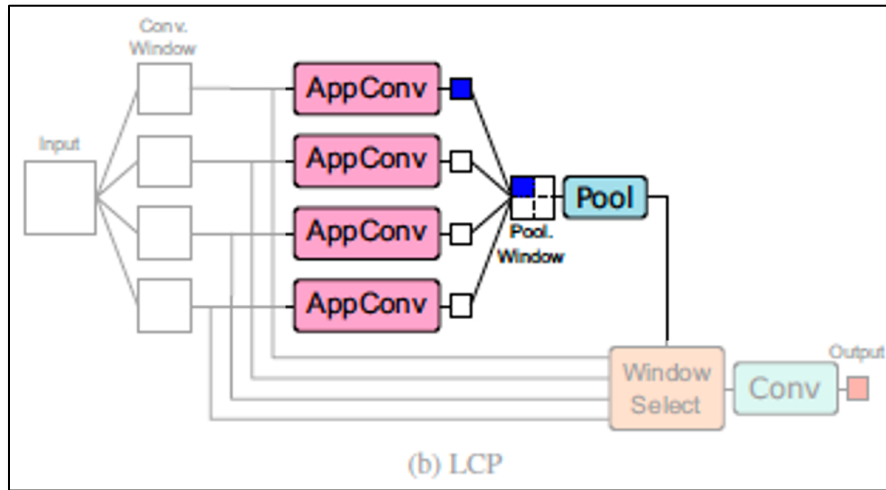
maximum value is present among the second set of values; and”

134. *Ujiie*’s modified system teaches the one or more non-transitory computer-readable storage media of claim 2. *Supra*, Claim 2.

135. *Ujiie*’s modified system stores instructions executed by its processor that perform a second iteration that computes a second set of AppConv values based on a second set of MSBs for each of the plurality of data sets. *Supra*, Claim 2. *Ujiie*’s modified system further examines a set of values computed for the plurality of data sets in the first iteration to determine whether a maximum value is present among the set of values. *Supra*, Claim 1(b). A POSITA would have understood that *Ujiie*’s modified system would analyze the second set of values computed in the second iteration to determine whether a maximum value is present in the same way as it analyzes the first set of values.

136. As described above, *Ujiie* utilizes “max pooling, which selects the max value from the pooling window.” *Ujiie*, 2. After computing the approximated convolution as modified by *Moons*’s precision scaling, *Ujiie* performs the “max pooling” operation on the results of the approximated convolutions. *Id.*, 4, Fig. 2b. Then, the maximum value is selected from the “pooling window[,]” and the approximate convolution that contains the maximum value is adopted. *Id.*, 3 (“...we adopt the approximation only to predict the **region to be selected in the max pooling**”). This process of selecting the max value from the pooling window is

illustrated in Figure 2b below, where maximum value is represented by the dark blue pixel:



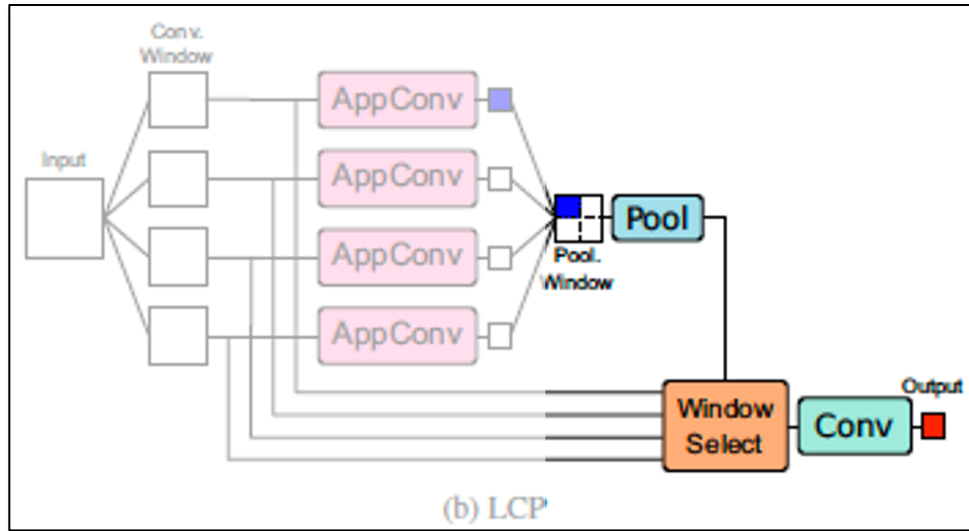
Id., Fig. 2b (modified for emphasis). Thus, *Ujiie*'s modified system would analyze the second set of approximated convolutions to determine whether the maximum value is present among the second set of values (i.e., *determine[s] whether the maximum value is present among the second set of values*) in the same way it analyzes the first set of approximated convolutions. *Supra*, Claim 1(b).

- b. Claim 3(b): “responsive to identifying the maximum value among the second set of values, performing, by the at least one processor, the full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value in the second iteration.”**

137. As stated above, *Ujiie-Moons-Kaul* (Ground 2) or alternatively, *Ujiie-Moons-Kaul-Ishii* teaches performing a full precision computation on the dataset that exhibited the maximum value, including in a second iteration with increased bit-

precision (i.e., “responsive to identifying the maximum value among the second set of values, performing, by the at least one processor, the full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value in the second iteration”). *Supra*, Claim 2; *see also* Claim 1(c) (describing similar process for the first iteration). A POSITA would have understood, as explained above, that *Ujiie*’s modified system would perform a full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value in the second iteration in the same way that it performs a full precision computation for the first iteration.

138. As described above, after *Ujiie*’s computing method completes Steps 1 and 2, *Ujiie* limits the full precision computation to the maximum value that was identified in the “max pooling” operation. *Ujiie*, 4 (“On the basis of this observation, we construct LCP framework with two main parts: ... (2) **a single regular (and exact) convolutional unit (Conv) to perform convolution only on the predicted window**”). Step 3 is described as “[c]omput[ing] Conv only to the input corresponding to the specified index in Step 2.” *Id.* In other words, the full precision computation is performed on the maximum value that was identified in Step 2 of *Ujiie*’s computing method. *Id.* This is demonstrated below in Figure 2b:



Id., Fig. 2b (modified for emphasis). Thus, *Ujiie*'s modified system would perform a full precision computation of the value that exhibited the maximum value in the second iteration (i.e., *performing the full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value in the second iteration*) in the same way that it performs a full precision computation for the first iteration if a maximum is identified. *Supra*, Claim 1(c).

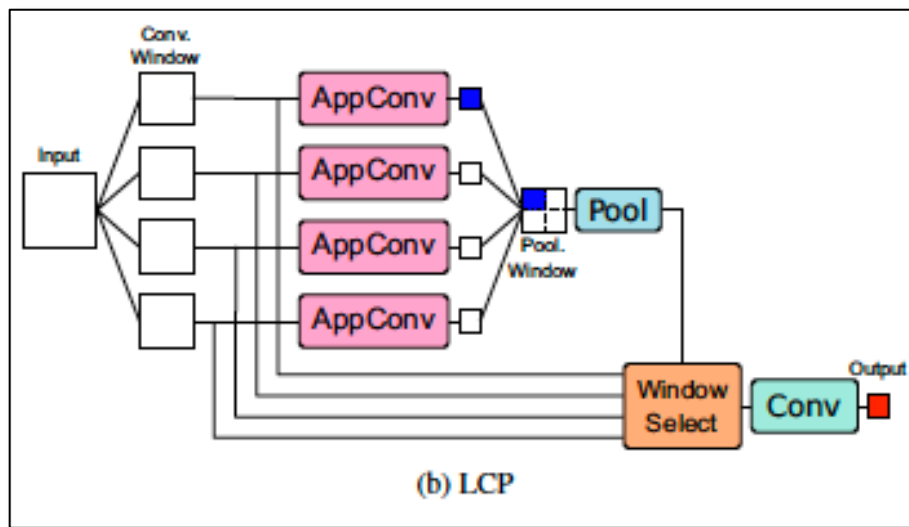
5. **Claim 4: The one or more non-transitory computer-readable storage media of claim 2, wherein the computing in each of the first iteration and the second iteration employs a convolution and a pooling.**

139. *Ujiie*'s modified system teaches the one or more non-transitory computer-readable storage media of claim 2. *Supra*, Claim 2.

140. *Ujiie* teaches performing convolution (e.g., approximate convolution "AppConv" and/or exact convolution "Conv") and pooling (i.e., max pooling) in both the first iteration and the second iteration. *Supra*, Claims 1(a)-(c), 2, 3; *see also*,

Ujii, 4 (“Figure 2b illustrates the LCP-based convolution and pooling operations.”),

Figure 2b:



Id., Fig. 2b.

6. **Claim 5: The one or more non-transitory computer-readable storage media of claim 4, wherein the convolution is a $N \times N$ convolution, where N is any integer.**

141. *Ujii*’s modified system teaches the one or more non-transitory computer-readable storage media of claim 4. *Supra*, Claim 4.

142. *Ujii* teaches specific network parameters, such as the size of its convolution filters, used for its LCP process on different datasets, including MNIST and CIFAR-10. *Ujii*, 5, Tables 1-2. In these network parameters, *Ujii* specifies that it uses a “**Convolution**” “kernel” that is “**5x5**” where N in this case is equal to 5. *Id.* Tables 1 and 2 showing the 5x5 convolutions are shown below:

Layer name	Parameter
Input image	size: 28x28, channel: 1
Convolution	kernel: 5x5, channel: 20
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 50
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 500
ReLU	
Fully connected	channel: 10
Softmax	

Table 1: Network parameters of CNN for MNIST dataset.

Layer name	Parameter
Input image	size: 32x32, channel: 3
Convolution	kernel: 5x5, channel: 64, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 128, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 256, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 512
ReLU	
Fully connected	channel: 10
Softmax	

Table 2: Network parameters of CNN for CIFAR-10 dataset.

Id., Tables 1 and 2 (annotated).

7. **Claim 6: The one or more non-transitory computer-readable storage media of claim 4, wherein the pooling is a NxN pooling, where N is any integer.**

143. *Ujii*e’s modified system teaches the one or more non-transitory computer-readable storage media of claim 4. *Supra*, Claim 4.

144. *Ujii*e teaches other specific network parameters, such as the size of the max pooling window, used for its LCP process on different datasets, including MNIST and CIFAR-10. *Ujii*e, 5, Tables 1-2. In these network parameters, *Ujii*e specifies that it uses a “**Max pooling**” “kernel” that is “**2x2**” where N in this case is equal to 2. *Id.* Tables 1 and 2 showing the 2x2 max pooling windows are shown below:

Layer name	Parameter
Input image	size: 28x28, channel: 1
Convolution	kernel: 5x5, channel: 20
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 50
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 500
ReLU	
Fully connected	channel: 10
Softmax	

Table 1: Network parameters of CNN for MNIST dataset.

Layer name	Parameter
Input image	size: 32x32, channel: 3
Convolution	kernel: 5x5, channel: 64, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 128, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 256, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 512
ReLU	
Fully connected	channel: 10
Softmax	

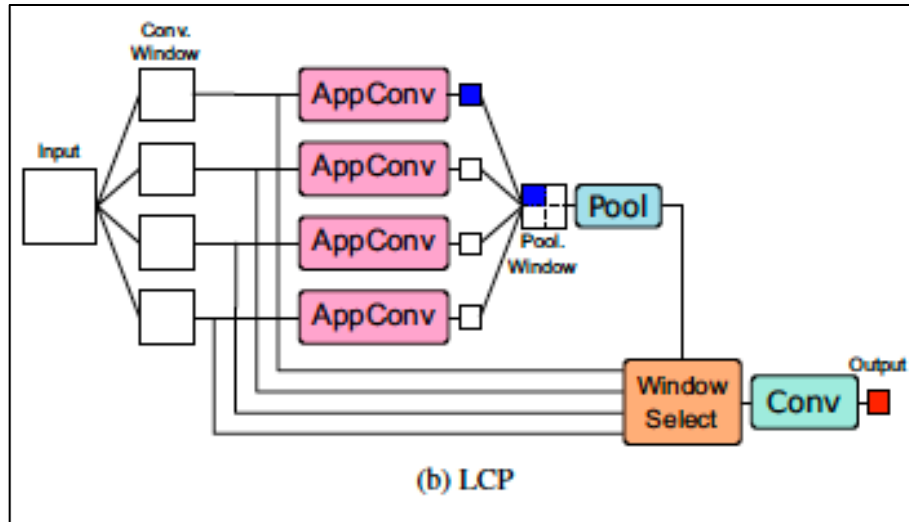
Table 2: Network parameters of CNN for CIFAR-10 dataset.

Id., Tables 1 and 2 (annotated).

8. Claim 8: The one or more non-transitory computer-readable storage media of claim 2, wherein at least one of the first iteration and the second iteration is performed with a precision less than that of the full precision computation.

145. *Ujii*’s modified system teaches the one or more non-transitory computer-readable storage media of claim 2. *Supra*, Claim 2.

146. *Ujii* teaches performing an approximate convolution “AppConv” followed by an exact full precision convolution “Conv.” *Supra*, Claims 1(a)-(c); *see also, Ujii*, 4 (teaching the two main parts of LCP are “(1) approximated convolution units (AppConv) to predict the feature selection in the pooling layer, and (2) a single regular (and exact) convolution unit (Conv) to perform convolution only on the predicted window”). This process is reproduced below in Figure 2b:

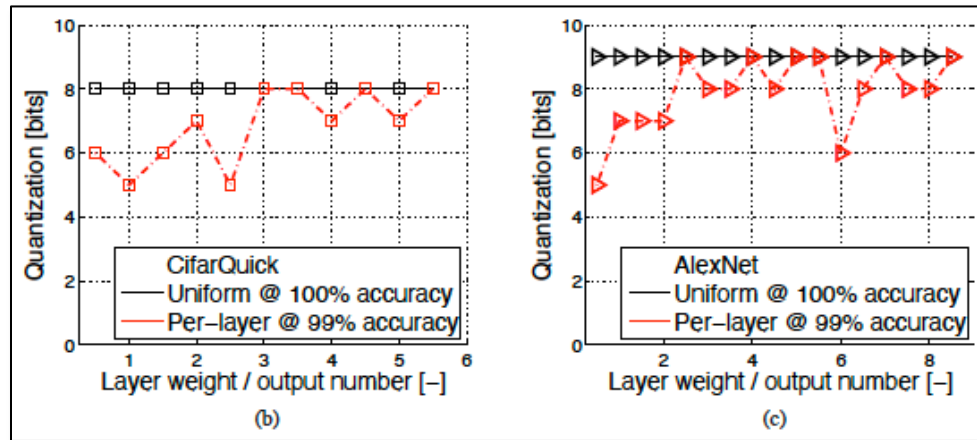


Id., Fig. 2b. As described above, it would have been obvious, and a POSITA would have been motivated to modify *Ujiie's* AppConv computations that use Sign Connect to approximate convolutions with *Moons's* precision scaling to perform those approximations. *Supra*, Claims 1(a)-(c). *Moons's* precision scaling approximates convolutions using less bits than the exact convolution. *Id.* *Ujiie's* modified approximated convolution is thus performed with a precision less than that of the full precision computation. Since the approximated computation is performed before the exact, full precision computation, at least the first iteration computation is performed at a lower precision than the full precision computation.

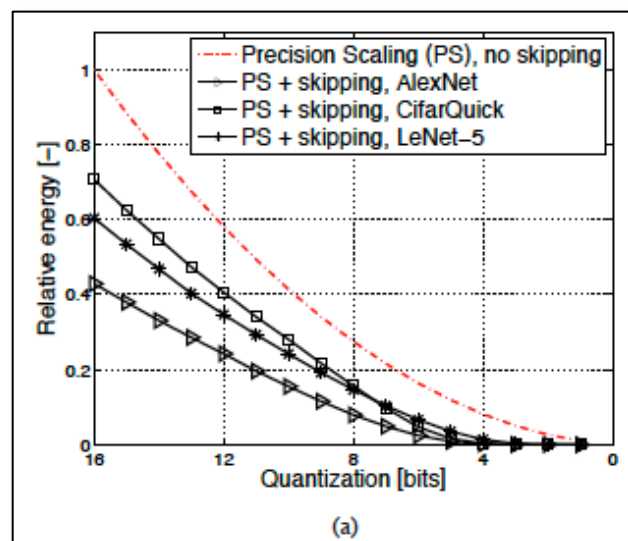
9. Claim 9: The one or more non-transitory computer-readable storage media of claim 8, wherein the precision is 8-bit precision.

147. *Ujiie's* modified system teaches the one or more non-transitory computer-readable storage media of claim 8. *Supra*, Claim 8.

148. *Moons* teaches its per-layer quantization uses 8-bit precision in at least some of its layers. For example, when using the CIFAR and AlexNet databases, multiple layers of convolution and pooling include 8-bit precision. *Moons*, 5, Figure 4:

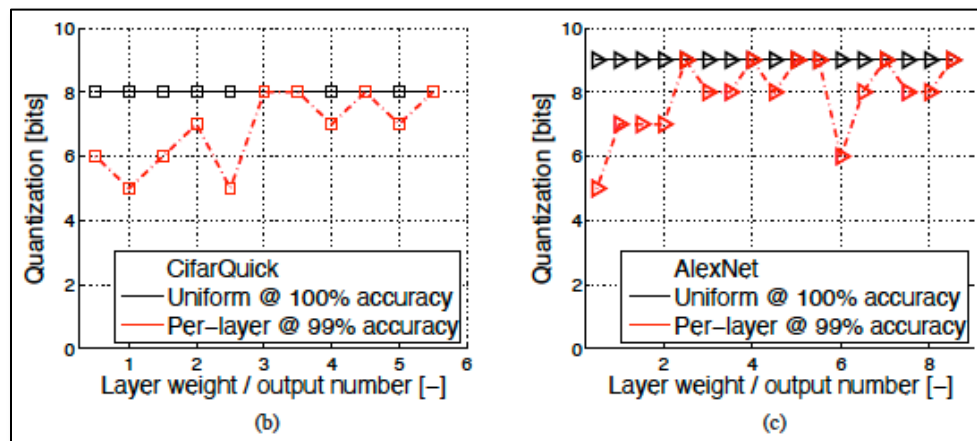


Id., Fig. 4. Furthermore, *Moons* teaches precision scaling at 8-bit precision decreases relative energy consumption by roughly 70% as compared to 16-bit precision. *Moons*, Figure 6a:



Id., Fig. 6a.

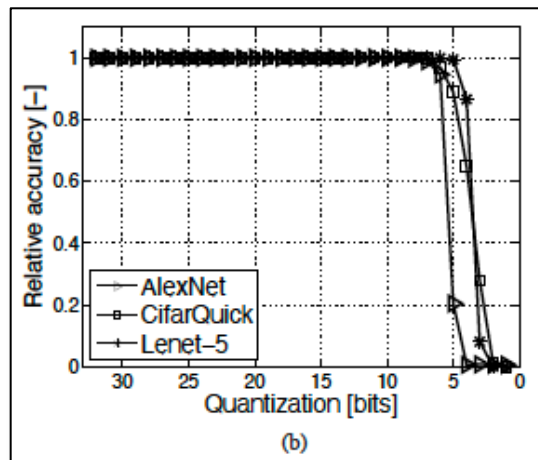
149. A POSITA would have been motivated and found it obvious to modify *Ujiie* to perform its approximate computations using 8-bit precision, as taught by *Moons*, for the same reasons stated above. *Supra*, Claim 1(a). For example, a POSITA would have recognized that, at least in some instances, using 8-bit precision for the approximate convolution would have provided the necessary accuracy for determining the maximum approximate value while still producing significant energy consumption savings. The achievement of necessary accuracy while maintaining energy consumption savings is demonstrated by *Moons*, specifically in Figures 4b and 4c:



Id., Fig. 4b and 4c. The figures above show that 8-bit width precision was needed in the higher layers of the network to achieve a target accuracy of 99%. *Id.* Generally, 8-bit width is required to maintain 99% accuracy starting at layer weight/output number 3 for the CIFAR database and approximately 2.5 for the AlexNet database.

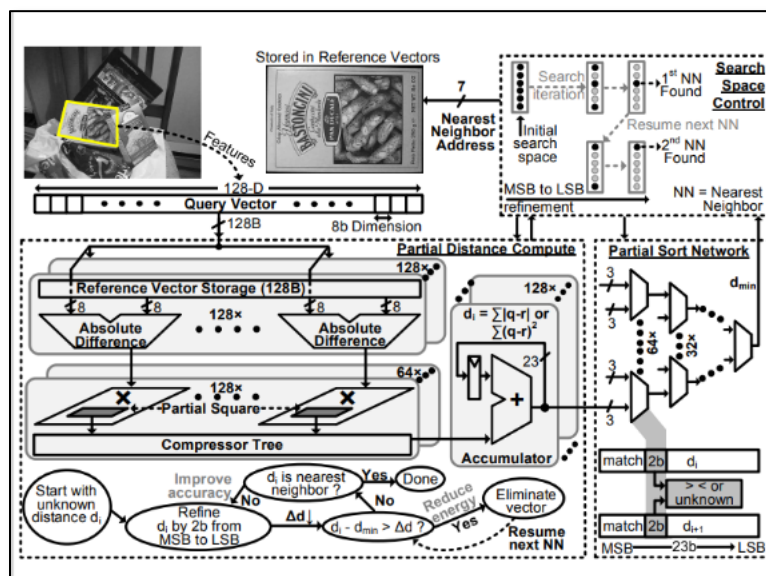
Moons acknowledges this phenomenon and states, “...there is a general trend of needing less bits in the lower layers of the network than in the higher layers.” *Id.*, 4.

150. Indeed, *Moons* teaches that “[e]ach network performs at its original accuracy up to quantization with no more than 8 bits. After that, the accuracy again drops quickly.” *Id.*, 4, Figure 2b:



Id., Fig. 2b. As such, a POSITA would have understood the benefits of using 8-bit precision for the approximate convolution as this would have provided significant energy savings, an express goal of *Ujiie*, while retaining sufficient precision to determine a maximum value and subsequently perform full precision computation on the maximum value’s data set. 8-bit precision is shown to be one of the lowest quantization levels that still achieves the original relative accuracy of each network. *Moons*, 4, Figure 2b. Utilizing the lowest possible precision to achieve the highest possible accuracy for the network would provide energy savings by creating a highly efficient network computation. *Id.*

151. A POSITA would have further understood that, if using a lower bit precision in the first iteration (e.g., 6-bit precision), the second iteration would have used 8-bit precision computation as taught by *Kaul* because *Kaul* teaches MSB- to LSB- based selective distance refinement using successive 2 bits in each iteration. *Kaul* iteratively increases the bit width when performing computations. *See e.g., Kaul*, 1 (“**Iterative refinement** of accumulated distance is interleaved as needed with **successive sort operations on 2b windows** from MSB to LSB”), *see also id.*, Figure 14.4.1; *supra*, Claim 2. For example, when performing a second iteration because the first 6-bit iteration did not identify a maximum value, *Ujiie-Moons* as modified by *Kaul* would have added 2 bits to the original 6-bit width, computing approximate convolutions on the 8-bit values. Specifically, *Kaul* teaches iteratively increasing precision by adding 2 bits when the nearest neighbor is not identified. *Kaul*, Figure 14.4.1; *Supra*, §VII.A.3.a.i.



Kaul, Fig. 14.4.1. Increasing the bit width increases the precision of the convolutional computation. Therefore, if a maximum is not found in the first 6-bit iteration, *Ujiie-Moons* as modified by *Kaul* would increase the precision in its second iteration by adding 2 bits to the values for approximate convolution. *Kaul*, 1. *Supra*, §VII.A.3.a.i. As such, a POSITA would have understood that *Ujiie*’s second iteration would be performed in 8-bit precision, as taught by *Moons* and *Kaul*. Namely, by adding 2 bits to the original 6-bit iteration, the second iteration would have a precision of 8-bits. The increase in precision from the first 6-bit iteration to the second 8-bit iteration would potentially allow for the identification of a maximum value in the performed computation. *Supra*, §VII.A.3.a.i.

VIII. OPINION: GROUNDS 4 AND 5

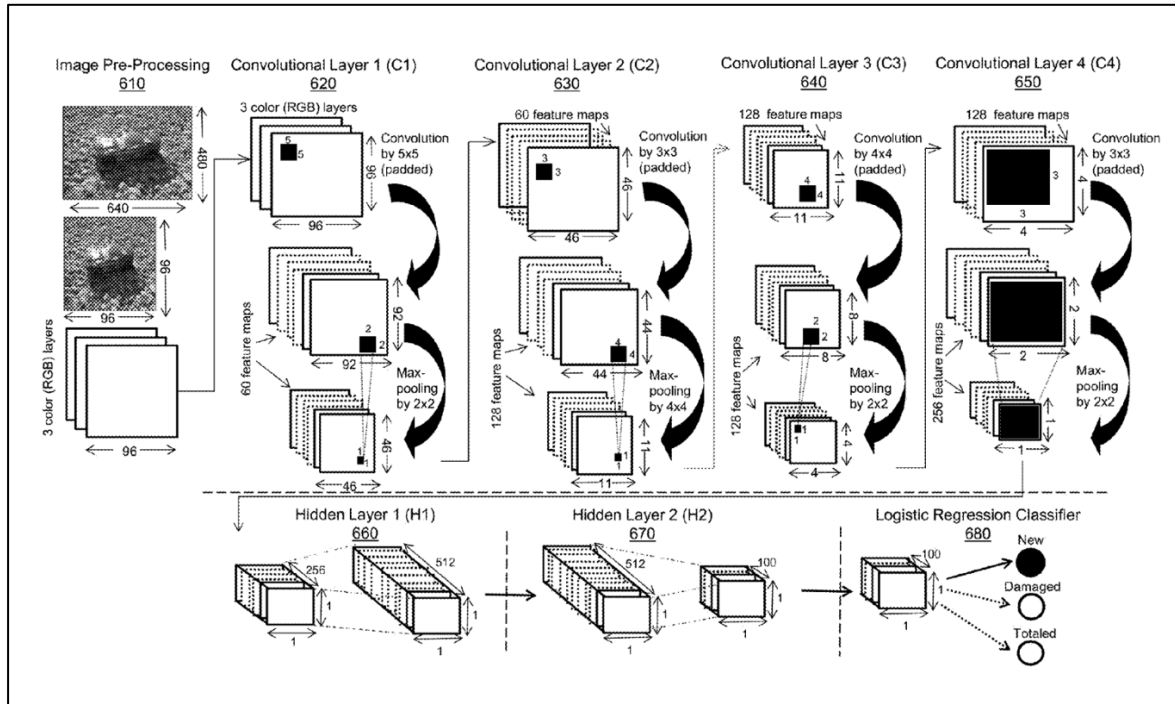
A. Obviousness over Ujiie, Moons, Kaul, and Ravindran, or Alternatively, Obvious Over Ujiie, Moons, Kaul, Ishii, and Ravindran

1. Overview of Ravindran

152. U.S. Patent Application Publication No. 2016/0259994 to Ravindran et al. (“*Ravindran*”) was filed on March 4, 2015, and published on September 8, 2016.

153. *Ravindran* is generally directed to “[a]n image processing system...[that] builds and trains an ensemble of deep learning models, such as convolutional neural networks (CNNs), to accurately and automatically perform image processing to detect particular attributes of objects in a digital image[.]”

Ravindran (EX1008), [0010]. In its method, *Ravindran* discloses various sizes of convolution and max pooling kernels. *Id.*, [0047]-[0050], Fig. 6.



Id., Fig. 6. Specifically, *Ravindran* teaches utilizing both 5x5 and 3x3 convolutional kernels, and 2x2 max pooling. *Id.*

2. Claim 7: The one or more non-transitory computer-readable storage media of claim 4, wherein the convolution is a 3x3 convolution, and the pooling is a 2x2 pooling.

154. *Ujii*'s modified system teaches the one or more non-transitory computer-readable storage media of claim 4. *Supra*, Claim 4.

155. As above, *Ujii* teaches performing "2x2" "Max pooling." *Ujii*, Tables 1-2; *supra*, Claim 6. In its examples, *Ujii* teaches using a 5x5 convolution kernel. *Ujii*, Tables 1-2; *supra*, Claim 5.

Layer name	Parameter
Input image	size: 28x28, channel: 1
Convolution	kernel: 5x5, channel: 20
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 50
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 500
ReLU	
Fully connected	channel: 10
Softmax	

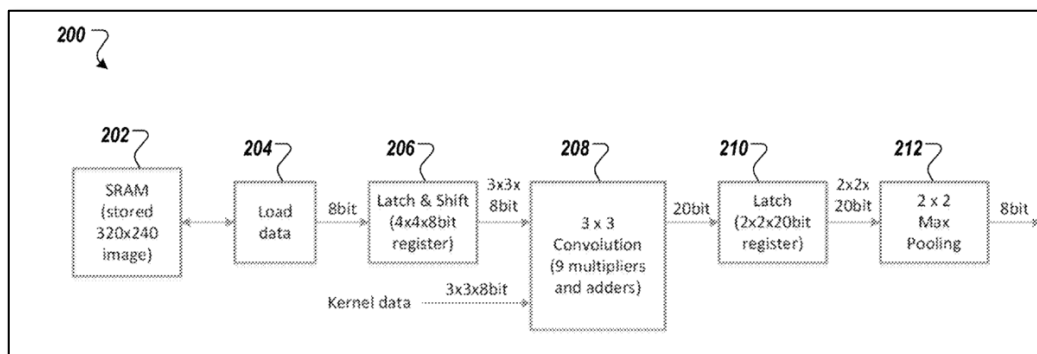
Table 1: Network parameters of CNN for MNIST dataset.

Layer name	Parameter
Input image	size: 32x32, channel: 3
Convolution	kernel: 5x5, channel: 64, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 128, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 256, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 512
ReLU	
Fully connected	channel: 10
Softmax	

Table 2: Network parameters of CNN for CIFAR-10 dataset.

Id., Tables 1 (top) and 2 (bottom). *Ujiie* does not expressly discuss examples using 3x3 convolution kernels. However, varying convolution size, including 3x3

convolutions, was an extremely well-known process before the invention date of the '831 Patent. For example, *Moons* explains that the size of convolution kernels “are typically very small on the order of 11 x 11, 5 x 5, or **3x 3.**” *Moons*, 6; *see also*, *Courbariaux*, 7 (“When using a ConvNet architecture with binary weights, the number of unique filters is bounded by the filter size. For example, in our implementation **we use filters of size 3 x 3**”). Indeed, even the inventors of the '831 Patent describe a prior art CNN including a 3x3 convolution and 2x2 max pooling. '831 Patent, 5:64-6:1 (“FIG. 2 shows a flow diagram, or block diagram, of a **conventional convolution and pooling scheme that employs 3x3 convolution, 2x2 pooling**”), Fig. 2.



Id., Fig 2. It would have been obvious to implement a 3x3 convolution kernel in *Ujiie*’s system, as taught by *Ravindran* for the reasons explained below.

a. Motivation to Combine Ujiie-Moons-Kaul-Ishii and Ravindran

156. A POSITA would have been motivated and found it obvious to modify *Ujiie*’s modified system to use a 3x3 convolutional filter, as taught by *Ravindran*.

For example, a POSITA would have understood that a 3x3 filter includes significantly fewer parameters (i.e., 9 parameters) than the 5x5 filter (i.e., 25 parameters) expressly taught by *Ujiie*. See also, *Ujiie*, Tables 1-2 (teaching using 5x5 filters on various datasets).

Layer name	Parameter
Input image	size: 28x28, channel: 1
Convolution	kernel: 5x5, channel: 20
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 50
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 500
ReLU	
Fully connected	channel: 10
Softmax	

Table 1: Network parameters of CNN for MNIST dataset.

Layer name	Parameter
Input image	size: 32x32, channel: 3
Convolution	kernel: 5x5, channel: 64, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 128, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 256, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Convolution	kernel: 5x5, channel: 512, padding: 2
Max pooling	kernel: 2x2, stride: 2
ReLU	
Fully connected	channel: 512
ReLU	
Fully connected	channel: 10
Softmax	

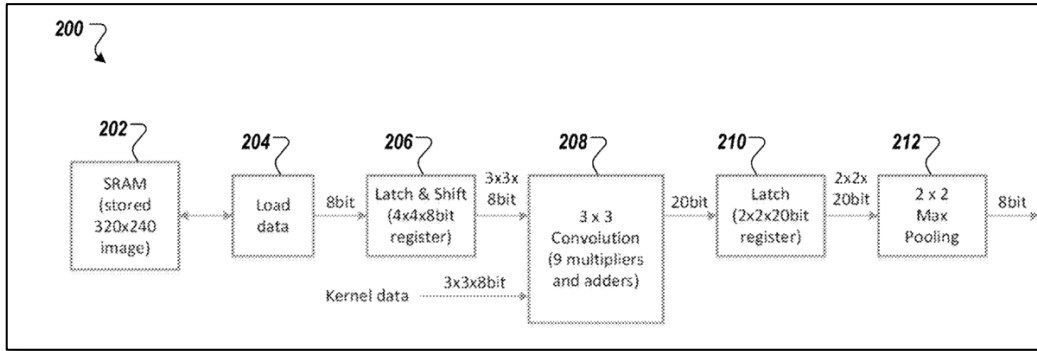
Table 2: Network parameters of CNN for CIFAR-10 dataset.

Id., Tables 1 (top) and 2 (bottom). Decreasing the number of parameters would have provided a lighter model that is easier to train given the need to calculate/optimize 9 instead of 25 parameters. During training, the CNN must backpropagate gradients to update each weight based on the error calculation identified at the classification

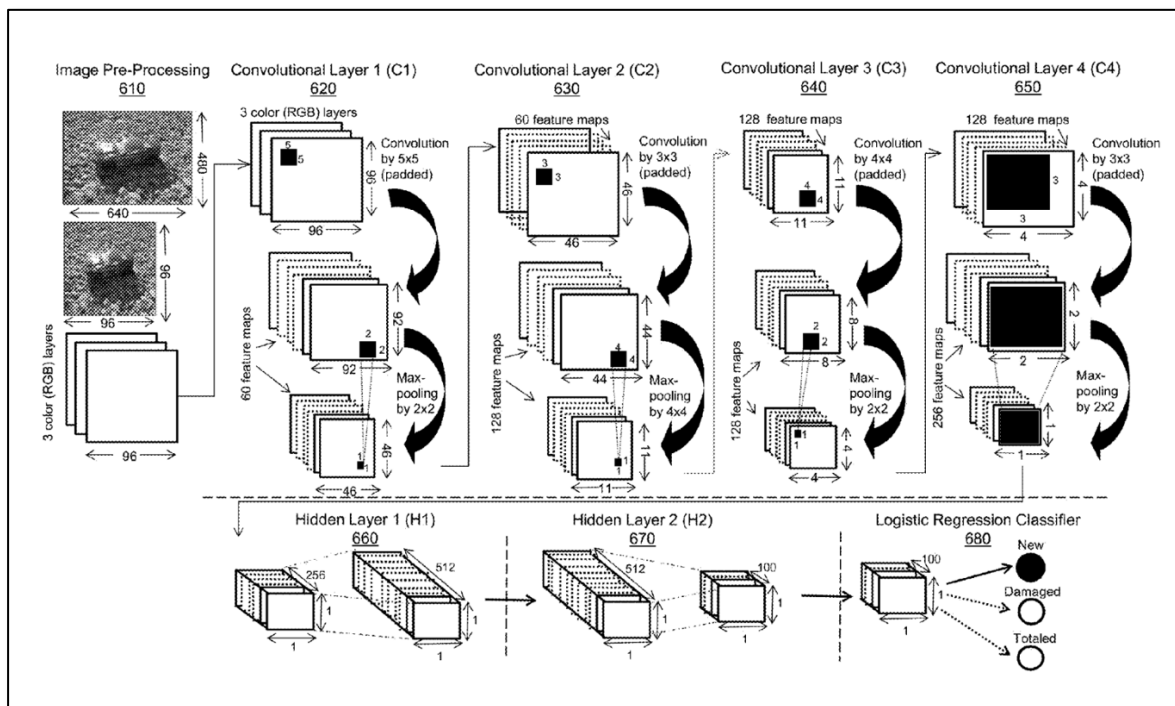
step. As such, this backpropagation and training step is faster and easier to train if there are only 9 parameters to update for each kernel as compared to a 5x5 kernel with 25 parameters to update each backpropagation step.

157. Moreover, this would have increased computational efficiency (i.e., decreased energy consumption), furthering an express goal of *Ujiie*, by providing decreased multiplication and addition steps per filter. A 3x3 convolution process only requires 9 total multiplication steps (i.e., 9 weight values multiplied by 9 input data values) followed by summations of those 9 values. On the other hand, a 5x5 convolution process is more computationally expensive since it requires 25 separate multiplication steps (i.e., 25 weight values multiplied by 25 input data values) followed by summations of those 25 values. Overall, a POSITA would have realized that the computational benefits of using a 3x3 convolutional filter would have furthered *Ujiie*'s goal of "reducing power consumption of CNN processors." *Ujiie*, 3.

158. A POSITA would have appreciated a reasonable expectation of success replacing *Ujiie*'s 5x5 convolutional filters with *Ravindran*'s 3x3 filters as this would have required simple programming techniques well within the skill of a POSITA. As mentioned above, the inventors of the '831 Patent admit that 3x3 convolutional filters were commonly used in the conventional CNN. '831 Patent, 5:64-6:1, Fig. 2.



Id., Fig. 2. *Moons* confirms the same. *Moons*, 6. It would have been a routine process to adjust *Ujii*'s convolutional filter to implement other well-known filter sizes. Furthermore, *Ujii*'s system already utilized a convolutional filter that was 5x5. *Ujii*, Tables 1-2. And as taught by *Ravindran*, a single system may even switch between 5x5 convolutional filters and 3x3 convolutional filters, indicating this change in filters was commonplace. *Ravindran*, [0047]-[0050], Fig. 6.



Id., Fig. 6. Indeed, such a process of switching between convolutional filter sizes was a common technique employed within the same CNN between layers. As such, modifying *Ujiie*'s convolution method using a 5x5 filter to use a 3x3 filter would have required simple programming techniques well known to a POSITA and resulted in an increase in computational efficiency both during training and usage of the system.

IX. CONCLUSION

159. I declare that all statements made herein of my knowledge are true, and that all statements made on information and belief are believed to be true, and that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Dated: September 26, 2025

By: 

Dr. Eli Saber

Appendix A

Prof. Eli Saber

Mobile: (585) 727-3126, Email: esaber228@gmail.com, HTTP: <http://people.rit.edu/essee>

EXPERIENCE

Computer Vision, Machine Learning, Digital Signal/Image/Video Processing, Video Compression and Coding

EDUCATION

Ph.D., Electrical Engineering, University of Rochester, Rochester, New York (March 1996)

Concentration: Signal/Image/Video Processing, Pattern Recognition, and Computer Vision.

Dissertation: Automatic image annotation and query-by-example using color, shape and texture information.

M.S., Electrical Engineering, University of Rochester, Rochester, New York (May 1992)

Concentration: Signal/Image/Video Processing, Pattern Recognition, Computer Vision, Communications.

B.S., Electrical and Computer Engineering, Summa Cum Laude, State University of New York at Buffalo, Buffalo, New York (May 1988).

Concentration: Computers, Microprocessors, Communications, Instrumentation.

A.S., Engineering Science, Mohawk Valley Community College, Utica, New York (May 1986)

CAREER MAJOR HIGHLIGHTS

❖ Academic Experience:

- **Teaching**: Taught/continue to teach undergraduate and graduate courses in the signal/image/video processing, computer vision, controls and communications focus areas with excellent feedback from students from 1996 to Present. Founded new coursework in Digital Video Processing and Pattern Recognition.
- **Research**: Conducted research on multiple acquired government and industry grants in the areas of: 1) image/video processing for multimedia, military and biomedical applications, 2) deep learning/artificial intelligence for object detection and tracking purposes, 3) three-dimensional scene reconstruction for remote sensing, and 4) color image processing for printing and multimedia applications. Graduated several PhD and MS students. Currently co-advising 2 PhD students.
- **External Funding**: Acquired external funding as PI in excess of \$3.5 Million and as PI/Co-PI in excess of \$5 Million over the period of September 2004 to present from government agencies and industry partners.
- **Publications**: 1 Book, 2 Special Issues, 38 Journal, 100 Conference and 11 Patents/Patent Publications.
- **Service**: Served/chaired several department, college and university level committees. Chaired the Graduate Program at the EE department for a period of five years. Served on the academic senate for a period of 9 years and on the Academic Senate Executive Committee for a period of 3 years.

❖ Industrial Experience: Worked for Xerox Corporation from 1988 until 2004 in a variety of engineering, managerial and scientific positions ending as Product Development Scientist and Manager.

❖ Consulting/Litigation Experience: Served as an expert in several patent and non-patent cases representing numerous companies. Provided declarations in support of multiple Inter Partes Reviews (IPR), expert opinions and reports with regards to validity/invalidity, infringement/non-infringement and domestic industry, tutorial at Markman hearing, multiple depositions, and multiple testimonies in court proceedings.

- ❖ **Professional Activities:** Served on several IEEE and SPIE conferences in various chair positions including Finance, Tutorials, Plenaries and Technical Program. Served as general chair for two conferences.
- ❖ **Honors & Awards:** RIT Trustees Award (2012), KGCOE Scholarship Award (2012), EME Gleason Professor (2011-2013), and PI Millionaire (2011).

RESEARCH INTERESTS

- Machine Learning for Imaging, Video and Remote Sensing Applications
- Image and Video Processing for Multimedia, Military & Biomedical Applications.
- Computer Vision and Three-Dimensional Scene Reconstruction.
- Color Image Processing for Printing and Multimedia Applications.

ACADEMIC EXPERIENCE

Professor (Associate 2004 – 2010, Full 2010 – Present), Department of Electrical and Microelectronic Engineering (EME), Kate Gleason College of Engineering (KGCOE), Rochester Institute of Technology.

Extended Faculty (2004 – Present), Center for Imaging Science (CIS), Rochester Institute of Technology.

Director of the Image, Video and Computer Vision Laboratory (2004 – Present)

Graduate Program Director (2010 – 2014)

- ❖ **Teaching:** Responsible for teaching undergraduate & graduate courses in Digital Signal Processing, Digital Image Processing, Digital Video Processing, Advanced Neural Network and Deep Learning, Engineering Analysis, Advanced Engineering Mathematics, Random Signal & Noise, Pattern Recognition, Communication Systems, Digital Data Communications, Computer Vision, Modern Control Theory, Linear Systems, and Matrix Methods. Typically teaching two courses per semester. Have taught three per semester on occasion to fulfill department needs.
- ❖ **Research:** Director of the Image, Video and Computer Vision Laboratory. Currently co-advising 2 PhD and 2 MS students on deep learning/machine learning/artificial intelligence applications for object detection and tracking. Advised several PhD and MS students in image/video segmentation, hierarchical image decomposition, video mosaicking/3-dimensional scene reconstruction, image/video understanding, object tracking/recognition. All students are/were funded under various government or corporate grants.
- ❖ **Funding:** Acquired funding as PI in excess of \$3.5 Million and as PI/Co-PI in excess of \$5 Million over the period of September 2004 – Present from various government agencies and industrial partners.
- ❖ **Service (Department, College and University):**
 - ❖ Advised/currently advising several undergraduate and graduate students on curriculum issues.
 - ❖ Member of the Academic Senate Resource Allocation and Budget Committee, the University Compensation Committee and of the Kate Gleason College of Engineering Promotion Committee.
 - ❖ Former Member (KGCOE representative) of the Academic Senate (2007 – 2013 and 2014-2017) and of the Academic Senate Executive Committee (2009-2010, 2016-2017).
 - ❖ Served as Chair of the RIT Vision 2025 committee per request from RIT President Dr William Destler. Committee was commissioned, over the summer of 2009, to review 80+ ideas submitted from all colleges and provide a recommendation to upper administration.
 - ❖ Former Member and Chair of KGCOE graduate committee.
 - ❖ Former Member of Graduate Council.
 - ❖ Served on several committees for curriculum development, recognition, and faculty search.

❖ **Professional Activities:**

- ❖ Senior member of the Institute of Electrical and Electronic Engineers (IEEE) society.
- ❖ Former Member of the IEEE Industry Technical Committee on DSP.
- ❖ Former Member of the IEEE Image and Multidimensional Digital Signal Processing (IMDSP) technical Committee.
- ❖ Former Area Editor for the Journal of Electronic Imaging.
- ❖ Former member of the Imaging Science and technology (IS&T) society.
- ❖ International Conference on Image Processing (ICIP) 2002 Finance Chair, ICIP 2007 and ICIP 2009 Tutorial Chair, ICIP 2012 General Chair, ICASSP 2017 Technical Program Chair, ICIP 2021 Plenary Chair.
- ❖ Co-founder and General Chair of the Video Surveillance and Transportation Imaging Conference within the Electronic Imaging Symposium.

❖ **Honors and Awards:**

- ❖ Awarded the Prestigious Trustees Scholarship award – the highest award at RIT with regards to research recognition
- ❖ Elected in 2011 as PI Millionaire.
- ❖ Awarded the EME Gleason Professor for 3 years (2011-2013).
- ❖ Elected as the Kate Gleason College of Engineering Scholarship award winner.

Adjunct Faculty Member, Dept. of Electrical & Computer Engineering, University of Rochester. (09/96-07/04)

- ❖ Taught undergraduate & graduate courses in Digital Signal Processing, Digital Image Processing, Pattern Recognition/Advanced Image Processing, Detection/Estimation Theory, and Analog & Digital Communications.
- ❖ Advised and graduated 1 Ph.D. student in the areas of “Image Understanding” and “Database Content Indexing”.
- ❖ Co-advised 1 Ph.D. student in the areas of “Watermarking”.
- ❖ Advised and graduated 1 MS student in the area of “Color Rendering” and “Printer Characterization”.
- ❖ Served as a committee member on several doctoral dissertations.
- ❖ Served on PhD qualifying examinations for the Signal Processing and Communications Concentration.
- ❖ Sought and captured funding from the National Science Foundation for the development of an intelligent image database system. Proposal funded for 4 years under NSF Grant IIS – 9820721
- ❖ Sought and captured industrial funding from Xerox Corporation for Printer Color Characterization & Digital Front End Object Oriented Rendering.

Adjunct Faculty Member, Dept. of Electrical Engineering, Rochester Institute of Technology. (03/98-07/04)

- ❖ Taught undergraduate and graduate courses in Pattern Recognition, Digital Video Processing, Random Signal & Noise, Image and Video Compression, and Communications.
- ❖ Advised and graduated a master student in the area of “Texture Classification”.

INDUSTRIAL EXPERIENCE

Product Development Scientist & Manager, Print Engine Development Unit, Xerox Corporation. (10/98-08/04)

Major responsibilities included:

- ❖ Lead the Image Science, Analysis and Evaluation area (12-15 direct reports and ~\$2 Million budget).
- ❖ Lead the development of highlight color specifications for the Sorrento print engine.
- ❖ Lead the development of color characterization algorithms for the iGen3 print engine.
- ❖ Lead the image quality integration of two color front end for the iGen3 Product.
- ❖ Lead the development of ROS and LED based imaging systems and image path architectures for upcoming highlight & full color products.
- ❖ Lead the development of xerographic hardware/algorithms & imaging systems for the DP92C highlight color product. (Product launched 9/30/99 and follow-on launched 4/20/00)
- ❖ Lead the research and development of image quality metrics for various product platforms and their dissemination throughout the Print Engine Development Unit and Xerox Corporation.

- ❖ Collaborate with the Department of Electrical & Computer Engineering (Univ. of Rochester) & the Center for Electronic Imaging Systems.

Advanced Development Scientist and Manager, Print Cartridge Delivery Unit, Xerox Corporation. (2/97 – 9/98)

Major responsibilities included:

- ❖ Establish the Advanced Design Laboratory (an imaging/xerographics lab) and provide technical and managerial leadership for the Electrical, Imaging and Xerographics Dept.
- ❖ Perform image processing and xerographic hardware/software design and development for low/mid volume color copiers and printers for current and future programs.
- ❖ Perform technology development, modeling, and product design for upcoming Xerox color products, specifically image on paper and image on belt products..
- ❖ Lead the development of the xerographic module for a color intermediate belt transfer product with direct technical and management responsibilities.
- ❖ Collaborate with the Department of Electrical & Computer Engineering (Univ. of Rochester) & the Center for Electronic Imaging Systems.

Research and Development Scientist, Production Systems Group, Xerox Corporation. (1/96-1/97)

Major responsibilities included:

- ❖ Lead the design and development of color characterization/management and image quality algorithms and specifications for digital front ends destined to drive high quality, high speed color print engines.
- ❖ Integrate color management & image processing algorithms into the Raster Image Processing module.
- ❖ Participate in the design and development of a high speed raster image processing architecture.
- ❖ Benchmark developed algorithms against existing products & systems both internally and externally.
- ❖ Collaborate with the Department of Electrical & Computer Engineering of the University of Rochester and the Center for Electronic Imaging Systems.

Research and Development Engineer, Corporate Research, Xerox Corporation (8/93-12/95) & Department of Electrical & Computer Engineering, University of Rochester. (1/95-12/95). Major responsibilities included:

- ❖ Design and develop query by image content and query by example image annotation algorithms utilizing color, shape, texture and motion cues. System is able to perform query by keywords, color, shape, texture, and/or a combination of the above cues.
- ❖ Design and develop intelligent image segmentation algorithms. These algorithms are currently utilized in the query by image content and query by example systems described above.
- ❖ Design and develop face detection and facial feature extraction approaches.
- ❖ Design and develop color characterization/calibration and image quality algorithms for Digital Front Ends aimed at driving high speed / high quality print engines.

(Note: Image annotation/content analysis research was done in conjunction with the Department of Electrical & Computer Engineering and Center for Electronic Imaging Systems leading to the Ph.D.)

Electronic, Computer and Instrumentation Engineer, New Toner/Developer Facility Engineering, Xerox Corporation. (6/88-7/93). Major responsibilities included:

- ❖ Provide design, development, installation, startup, and training for multiple toner production facilities.
- ❖ Provide development and implementation of control system database, software and displays for several systems.
- ❖ Evaluate vendor supplied electrical specifications and drawings.
- ❖ Manage and coordinate the efforts of technicians, electrical support, construction crew, and industrial workforce during the design, construction, startup, and implementation phases.
- ❖ Supervise and complete a number of upgrade projects for toner & photoreceptor production including software development, preparation of electrical design, procurement of necessary equipment and parts, supervision of technicians, contractors and industrial workforce, and scheduling of construction.

During this time, I gained extensive experience in the following systems: Fisher distributive control, unit operation controller, Provue console, Acrison material handling, Werner and Pfleiderer extrusion, Alpine air grinding, Majac/Micropul centrifugal classifiers, dry/wet material screening, Waeschle and others bulk powder storage and pneumatic convey, Ingersoll Rand and Joy compressed air equipment, Statistical process control.

CONSULTING EXPERIENCE

Expert on Legal Cases: Served as an expert in several patent and non-patent cases representing numerous companies. Provided declarations in support of multiple Inter Partes Reviews (IPR), expert opinions and reports with regards to validity/invalidity, infringement/non-infringement and domestic industry, tutorial at Markman hearing, multiple depositions, and multiple testimonies in court proceedings.

Summary of Experience:

- IPR/Declarations: Provided declarations in support of multiple IPR proceedings.
- Depositions: Provided multiple depositions.
- Court Testimony: Testified in International Trade Commission Court in Sept. 2018, Oct. 2019 and Sept. 2024
- Expert Opinions: Provided multiple expert opinions and reports in support of validity/invalidity, infringement/non-infringement, domestic industry and rebuttals.
- Claim Construction: Provided reports in support of claim construction.
- Hearings: Attended one Markman hearing and provided a brief technology tutorial.

Industrial Training: Developed and conducted industrial training at Xerox Corporation for engineering personnel over three separate summer periods in the areas of signal and image processing, color engineering, and control systems with a distinct focus on digital front ends and print engines. I have also provided pattern recognition/shape matching expertise for Leica corporation.

BOOKS

1. S. Dianat and E. Saber, “Advanced Linear Algebra for Engineers with MATLAB”, CRC press, February 2009.

SPECIAL ISSUES

1. H. J. Trussell, E. Saber and M. Vrhel, “Color Image Processing”, IEEE SP magazine, January 2005.
2. R. Loce and E. Saber, “Video Surveillance and Transportation Imaging”, Journal of Electronic Imaging, 22(4), Dec. 2013

PEER-REVIEWED JOURNAL PUBLICATIONS

1. M. Sharma, J. Heard, E. Saber and P. Markopoulos, “Convolutional Neural Network Compression via Dynamic Parameter Rank Pruning”, IEEE access Journal, January 2025.
2. S. Singh, E. Saber, P. Markopoulos and J. Heard, “Regulating Modality Utilization within Multimodal Fusion Networks”, Multidisciplinary Digital Publishing Institute (MDPI) Sensors – Special Issue: Deep Learning Methods for Aerial Imagery, Sept. 2024
3. M. Sharma, M. Dhararaj, S. Karnam, D. Chachlakis, R. Ptucha, P. Markopoulos and E. Saber, “YOLOrs: Object Detection in Multimodal Remote Sensing Imagery”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 14, 2020.
4. S. Piramanayagam, E. Saber, and N. Cahill, “Gradient Driven Unsupervised Video Segmentation using Deep Learning Techniques”, Journal of Electronic Imaging, 29 (1), 013019, 2020
5. U. Gewali, S. Monteiro and E. Saber, “Spectral Super-Resolution with Optimized Bands”, Remote Sensing, 11(14), 2019

6. U. Gewali, S. Monteiro and E. Saber, "Gaussian Processes for Vegetation Parameter Estimation from Hyperspectral Data with Limited Ground Truth", *Journal of Remote Sensing*, 11(13), 2019.
7. Y. Liang, P. Markopoulos and E. Saber, "Spatial-Spectral Segmentation of Hyperspectral Images for Subpixel Target Detection", *Journal of Applied Remote Sensing*, 13(3), 2019.
8. Y. Liu, S. Piramanayagam, S. Monteiro, and E. Saber. "Semantic segmentation of multisensor remote sensing imagery with deep ConvNets and higher-order conditional random fields." *Journal of Applied Remote Sensing* 13, no. 1 (2019): 016501
9. U. Gewali, S. Monteiro and E. Saber, "Machine Learning Based Hyperspectral Image Analysis : A Survey", arXiv preprint, 2018.
10. S. Piramanayagam, E. Saber, W. Schwartzkopf, F.W. Koehler, "Supervised Classification of Multisensor Remotely Sensed Images using a Deep Learning Framework", *Remote Sensing Journal*, 10 (9), 2018.
11. S. R. Vantaram, Y. Hu, E. Saber and S. Dianat, "Synthesis of Intensity Gradient and Texture Information for Efficient Three-Dimensional Segmentation of Medical Volumes", *Journal of Medical Imaging* 2, no. 2 (2015): 024003-024003.
12. S. R. Vantaram, S. Piramanayagam, E. Saber and D. Messinger, "Spatial Segmentation of Multi/Hyperspectral Imagery by Fusion of Spectral-Gradient-Textural Attributes", *Journal of Applied and Remote Sensing*, Vol. 9, No. 1, pp. 095086 (1-37), 2015.
13. S. R. Vantaram and E. Saber, "A Survey of Contemporary Trends in Color Image Segmentation", *Journal of Electronic Imaging*, 21(4), 040901, Oct-Dec 2012.
14. M. S. Erkilinc, M. Jaber, E. Saber, "Text, Photo and Line Extraction in Scanned Documents", *Journal of Electronic Imaging*, Vol. 21, 033006, July 2012.
15. T. Keane, E. Saber, H. Rhody, A. Savakis and J. Raj, "Practical Image Registration Concerns Overcome by the Weighted and Filtered Mutual Information Metric", *Journal of Electronic Imaging*, Vol. 21(2), 023029, June 2012.
16. P. Gurram, E. Saber and H. Rhody, "Semi-automated System for three-dimensional Modeling of Buildings from Aerial Video", *Journal of Electronic Imaging*, Vol. 21(1), 013007, Jan-Mar 2012.
17. M. Jaber and E. Saber, "Probabilistic Approach for Extracting Regions of Interest in Digital Images", *Journal of Electronic Imaging*, Vol. 19, No. 2, April - June 2010.
18. X. Fan, H. Rhody and E. Saber, "A Spatial Feature Enhanced MMI Algorithm for Multimodal Airborne Image Registration", *IEEE Transaction on Geoscience and Remote Sensing*, Vol. 48, Issue 6, pp. 2580 – 2589, 2010.
19. P. Gurram, E. Saber and H. Rhody, "A Segment-Based Mesh design for Building Parallel-Perspective Stereo Mosaic", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 48, No. 3, March 2010.
20. S. Vantaram, E. Saber, S. Dianat, M. Shaw and R. Bashkar, "Multiresolution Adaptive and Progressive Gradient-based color image Segmentation", *Journal of Electronic Imaging*, Volume 19, Number 1, pp. 1-21, January-March 2010.
21. L. Garcia, E. Saber, S. Vantaram, V. Amuso, M. Shaw and R. Bhaskar, "Automatic Image Segmentation by Dynamic Region Growth and Multi-resolution Merging", *IEEE Transactions on Image Processing*, Vol. 18, No. 10, Oct. 2009.
22. H. Santos, E. Saber, and W. Wu, "Streak Detection in Mottled and Noisy Images", *Journal of Electronic Imaging*, Vol. 16, No. 4, 2007.
23. O. Ugbeme, E. Saber and W. Wu, "An Automated Algorithm for the Identification of Artifacts in Mottled and Noisy Images", *Journal of Electronic Imaging*, Vol. 16, No. 3, 2007.
24. V. Mistic, V. Sampath, Y. Yu and E. Saber, "Prostate Boundary Detection and Volume Estimation Using TRUS Images for Brachytherapy Applications", *International Journal of Computer Assisted Radiology and Surgery*, Vol. 2, No. 2, pp. 87-98, August 2007.
25. A. Ononye, A. Vodacek and E. Saber, "Towards Automatic Extraction of Fire Line Parameters from Multispectral Infrared Images", *Journal of Remote Sensing of Environment*, Vol. 108, pp. 179 – 188, 2007.
26. E. Saber, S. Dianat and L. Mestha, "DSP utilization in Digital Color Printing", *IEEE SP Magazine*, July 2005.
27. E. Saber, Y. Xu, and A. M. Tekalp, "Partial Shape Recognition by sub-matrix matching for partial matching guided image labeling", *Pattern Recognition*, Vol. 38, pp. 1560 – 1573, 2005.
28. M. Celik, G. Sharma, A. M. Tekalp, E. Saber, "Lossless Generalized-LSB Data Embedding", *IEEE Trans. on Image Processing*, Vol. 14, No. 2, pp. 253 – 266, February 2005.

29. M. Vrhel, E. Saber, and H. J. Trussell, "Color Image Generation and Display Technologies: An overview of methods, devices, and research", IEEE Signal Processing Magazine, January 2005.
30. H. J. Trussell, E. Saber and M. Vrhel, "Color Image Processing: Basics and Special Issue Overview", IEEE Signal Processing Magazine, January 2005.
31. Y. Xu, E. Saber, and A. M. Tekalp, "Dynamic Learning from Multiple Examples for Semantic Object Segmentation and Search", Computer Vision and Image Understanding, Vol. 95, No. 3, pp. 334-353, Sept 2004.
32. Y. Xu, P. Duygulu, E. Saber, A. M. Tekalp, and F. T. Yarman-Vural, "Object-Based Image Labeling through Learning by Example and Multi-Level Segmentation", Pattern Recognition, Vol. 36 (6), pp. 1407-1423, June 2003.
33. Y. Xu, E. Saber, and A. M. Tekalp, "Object Segmentation and Labeling by Learning from Examples", IEEE Transactions on Image Processing, Vol. 12, No. 6, June 2003.
34. M. Celik, G. Sharma, E. Saber, and A. M. Tekalp, "Hierarchical watermarking for secure image authentication with localization", IEEE Trans. on Image Processing, vol. 11, no. 6, June 2002.
35. M. Xia, E. Saber, G. Sharma, and A. M. Tekalp, "End-to-End Color Calibration by Total Least Squares Regression", IEEE Transactions on Image Processing, Vol. 8, No. 5, May 1999.
36. E. Saber and A. M. Tekalp, "Facial Pattern Detection and Eye Localization using Color, Shape and Symmetry-Based Cost Functions", Pattern Recognition letters, Vol. 19, 1998.
37. E. Saber and A. M. Tekalp, "Integration of Color, Shape and Texture for Automatic Image Classification, Annotation and Retrieval", Journal of Electronic Imaging, Vol. 7, No. 3, July 1998.
38. E. Saber and A. M. Tekalp "Region-Based Affine Shape Matching for Automatic Image Annotation and Query-by-Example", Journal of Visual Communication and Image Representation, March 1997.
39. E. Saber, A. M. Tekalp, and G. Bozdagi, "Fusion of Color and Edge Information for Improved Segmentation and Edge Linking,", Image and Vision Computing, Vol. 15, 1997.
40. E. Saber, A. M. Tekalp, R. Eschbach and K. Knox, "Automatic Image Annotation using Color Classification", Graphical Models and Image Processing, Volume 58, Number 2, March 1996.

CONFERENCE & WORKSHOP PUBLICATIONS

1. M. Sharma, J. Heard, E. Saber and P. Markopoulos, "VEDAI-R Dataset", in preparation for submission to SPIE.
2. D. Velychko, S. Singh, P. Markopoulos, E. Saber, J. Heard, "Image Preprocessing and YOLO Architectures for Enhanced Small and Slow-Moving Object detection", Western NY Image and Signal Processing Workshop, 2024
3. S. Singh, P. Markopoulos, E. Saber, J. Lew and J. Heard, "Measuring Modality Utilization in Multi-Modal Neural Networks", 2023 IEEE Conference on Artificial Intelligence (CAI 2023), Santa Clara, CA.
4. S. Singh, M. Sharma, J. Heard, J. Lew, E. Saber and P. Markopoulos, "Multimodal Aerial View Object Classification with Disjoint Unimodal Feature Extraction and Fully-Connected-Layer Fusion", SPIE: Defense and Commercial Sensing 2023, Orlando, FL.
5. M. Sharma, P. P. Markopoulos, E. Saber, M. S. Asif, and A. Prater-Bennette, "Convolutional Auto-Encoder with Tensor-Train Factorization," Proc. International Conference on Computer Vision, (ICCV 2021), RLS-CV workshop.
6. M. Sharma, P. P. Markopoulos, and E. Saber, "YOLOrs-LITE: A Lightweight CNN for Real-time Object Detection in Remote Sensing," Proc. IEEE International Geoscience and Remote Sensing Symposium (IEEE IGARSS), Brussels, Belgium, July 2021.
7. M. Dhanaraj, M. Sharma, T. Sarkar, S. Karnam, D. Chachlakis, R. Ptucha, P. Markopoulos and E. Saber, "Vehicle Detection from Multi-modal Aerial Imagery using YOLOv3 with Mid-level Fusion", Big Data II: Learning, Analytics, and Applications, SPIE, April 2020.
8. Y. Liu, S. Piramanayagan, S. Monteiro and E. Saber, "Dense Semantic Labeling of Very High Resolution Aerial Imagery and LIDAR with Fully Convolutional Neural Networks and Higher Order CRFs", CVPR, 2017
9. Y. Liu, S. Piramanayagan, S. Monteiro and E. Saber, "Semantic segmentation of remote sensing data using Gaussian processes and higher order CFRs", IGARSS, Fort Worth, TX, July 2017.

10. Y. Liang, S. T. Monteiro, and E. Saber, "Gaussian processes for object detection in high resolution remote sensing images", IEEE International Conference on Machine Learning and Applications (ICMLA 2016), Anaheim, CA, December 2016.
11. S. Piramanayagam, W. Schwartzkopf, F.W. Koehler, E. Saber, "Classification of remote sensed images using random forests and deep learning framework", Proc. SPIE 10004, Image and Signal Processing for Remote Sensing XXII, 100040L, October 2016;
12. Y. Liang, S. T. Monteiro, and E. Saber, "Transfer learning for high resolution aerial image classification", to appear in IEEE Applied Imagery Pattern Recognition Workshop (AIPR 2016), Washington, D.C., October 2016.
13. Y. Liang, P. P. Markopoulos, and E. Saber, "Subpixel target detection in hyperspectral images with local matched filtering in SLIC superpixels", IEEE Workshop on Hyperspectral Image and Signal Processing: Evolutions in Remote Sensing (WHISPERS 2016), Los Angeles, CA, August 2016.
14. Y. Hu, S. Monteiro and E. Saber, "Super Pixel Based Classification using Conditional Random Fields for Hyperspectral Images", ICIIP 2016, Phoenix, AZ.
15. O. de Lima, S. Janakiraman, E. Saber, D. C. Day, M. Shaw, P. Bauer, R. S. Twede, and P. Lea, "Signature Line Detection in Scanned Documents", ICIIP 2016, Phoenix, AZ.
16. Y. Liang, P.P. Markopoulos, and E. Saber, "Subpixel Target Detection in Hyperspectral Images from Superpixel Background Statistics", IGARSS, Beijing, China, July 2016.
17. Y. Liu, S.T. Monteiro, and E. Saber. "Vehicle detection from aerial color imagery and airborne LiDAR data", IGARSS, Beijing, China, 2016.
18. Y. Wang, J. Mathew, E. Saber, D. Larson, P. Bauer, G. Kerby and J. Wagner. "Scanned Document Enhancement Based on Fast Text Detection," International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 2016
19. Y. Liu, S. Monteiro, and E. Saber, "An Approach for Combining Airborne LiDAR and High Resolution Aerial Color Imagery using Gaussian Processes", Proc. SPIE 9643, Image and Signal Processing for Remote Sensing XXI, Oct. 2015.
20. Y. Liang, N. Cahill, E. Saber and D. Messinger, "A Game-Theoretic Tree Matching Approach for Object Detection in High Resolution Remotely Sensed Images", Proc. SPIE 9643, Image and Signal Processing for Remote Sensing XXI, Oct. 2015.
21. Y. Hu, S. T. Monteiro and E. Saber, "Comparing Inference Methods for Conditional Random Fields for Hyperspectral Image Classification", *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, Tokyo, Japan, 2015
22. Y. Hu, N. Cahill, S. Monteiro, E. Saber and D. Messinger, "Dimensionality Reduction for Hyperspectral Imagery Classification in Conditional Random Fields", Proc. SPIE 9643, Image and Signal Processing for Remote Sensing XXI, Oct. 2015.
23. S. Piramanayagam, P. J. Cutler, W. Schwartzkopf, F.W. Koehler, E. Saber, "Application of gradient based image segmentation to SAR imagery", IGARSS, Milan, Italy, July 2015.
24. S. Piramanayagam, E. Saber, N. D. Cahill, and D. Messinger, "Shot Boundary Detection and Label Propagation for Spatio-Temporal Video Segmentation", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Feb, 2015.
25. M. Yousefhusien, R. Easton, R. Ptucha, M. Shaw, B. Bradburn, J. Wagner, D. Larson and E. Saber, "Flatbed Scanner Simulation to Analyze the Effect of Detector's Size on Color Artifacts", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Feb. 2015.
26. K. Shah, E. Saber and K. Verrier, "Improved Metrology of Implant Lines on Static Images of Textured Silicon Wafers using Line Integral Method", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Feb. 2015.
27. Y. Hu, E. Saber, S. Monteiro, N. Cahill and D. Messinger, "Classification of Hyperspectral Images Based on Conditional Random Fields", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Feb. 2015.
28. R. Kothari, E. Saber, M. Nelson, M. Stauffer and D. Bohan, "Image Enhancement for Low Resolution Display Panels", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Feb. 2015.
29. Y. Liang, A. H. Syed, N. Cahill, E. Saber, and D. Messinger, "Application of Tree Matching Techniques to High Resolution Remotely Sensed Images toward Object Detection", GEOBIA, May 2014.
30. Y. Hu, A. H. Syed, E. Saber, N. Cahill, and D. Messinger, "Dynamic scale-space representation based on a MRF region merging model", GEOBIA, May 2014
31. Y. Liu, M. Helguerra and E. Saber, "Measurement of blood flow velocity in vivo video sequences with motion estimation methods", SPIE Medical Imaging, San Diego, 2014.

32. Y. Wang, O. F. de Lima, and E. Saber, "Improved edge directed Super resolution with hardware realization for surveillance, transportation, and multimedia applications", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Feb. 2014.
33. J. Whitesell, D. Patru, E. Saber, G. Roylance and B. Larson, "Design for Implementation of Color Image Processing Algorithms", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Feb. 2014
34. R. Toukatly, D. Patru, E. Saber, E. Peskin, G. Roylance and B. Larson, "Performance Analysis of a Color Space Conversion Engine Implemented using Dynamic Partial Reconfiguration", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Feb. 2013.
35. O. De Lima, S. R. Vantaram, S. Piramanayagam, E. Saber and K. Bengtson, "An Edge Directed Super Resolution Technique for Surveillance and Printing Applications", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Feb. 2013.
36. M. YousefHussien, K. Garvin, D. Dalecki, E. Saber and M. Helguera, "Three-Dimensional Volume Analysis of Vasculature in Engineered Tissues", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Feb. 2013
37. A. Mykyta, D. Patru, E. Saber. G. Roylance and B. Larson, "Reconfigurable Framework for High-Bandwidth Stream-Oriented Data Processing", IEEE International SoC Conference (SOCC), Buffalo, NY, Sept. 2012.
38. S. R. Vantaram, and E. Saber, "Unsupervised Video Segmentation by Dynamic Volume Growing and Multivariate Volume Merging using Color-Texture-Gradient Features", *IEEE International Conference on Image Processing, Orlando, Florida, September 2012*.
39. A. H. Syed, E. Saber and D. Messinger, "Encoding of Topological Information in Multi-Scale Remotely Sensed Data: Applications to Segmentation and Object-Based Image Analysis", International Conference on geographic Object-Based Image Analysis (GEOBIA), Rio de Janeiro, May 2012. (Best paper Award)
40. S. Piramanayagam, E. Saber and D. Heavner, "Measurement of Glucose by Image Processing of Thin Film Slides", SPIE: Medical Imaging, San Diego, CA, 2012.
41. Y. Hu, E. Saber, S. Dianat, S. R. Vantaram and V. Abhyankar, "An Automatic Approach for 3D Registration of CT Scans", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Jan. 2012
42. S. R. Vantaram, E. Saber, S. Dianat, Y. Hu and V. Abhyankar, "Semi-Automatic 3-D Segmentation of Computed Tomographic Imagery by Iterative Gradient-Driven Volume Growing", *IEEE International Conference on Image Processing, Brussels, Belgium, September 2011*.
43. S. R. Vantaram, E. Saber and D. Messinger, "An Intensity-Gradient-Texture Guided Methodology for Automatic Spatial Segmentation of Remotely Sensed Multi/Hyperspectral imagery", *IEEE International Conference on Image Processing, Brussels, Belgium, September 2011*.
44. S. R. Vantaram, and E. Saber, "A Method for Improved Localization of Edges in multi/hyperspectral Imagery", Proceedings of the SPIE Optical Engineering + Applications Conf., San Diego, CA, Aug. 2011.
45. M. Jaber, M. S. Bailly, Y. Wang, and E. Saber, "An image-set for identifying multiple regions/levels of interest in digital images," *Proceedings of the SPIE Optical Engineering + Applications Conference*, San Diego, CA, Aug. 2011.
46. M. S. Erkilinc, M. Jaber, E. Saber, P. Bauer and D. Depalov, "Page layout analysis and classification in complex scanned documents," *Proceedings of the SPIE Optical Engineering + Applications Conference*, San Diego, CA, Aug. 2011.
47. S. R. Vantaram, and E. Saber, "An Adaptive Bayesian Clustering and Multivariate Region Merging-based Technique for Efficient Segmentation of Color Images", IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, May 2011.
48. A. H. Syed, E. Saber and D. Messinger, "Scale-space Representation of Remote Sensing Images using an Object-Oriented Approach", SPIE Defense, Security, and Sensing, Orlando, Florida, Apr 2011.
49. M. Jaber and E. Saber, "Image Understanding Algorithm for Segmentation Evaluation and Region-of-Interest Identification using Bayesian Networks", SPIE Defense, Security, and Sensing, Orlando, Florida, Apr 2011.
50. T. Keane, E. Saber, H. Rhody, A. Savakis and J. Raj, "Unsupervised automated panorama creation for realistic surveillance scenes through weighted mutual information registration ", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Jan. 2011.
51. M. S. Erkilinc, M. Jaber, E. Saber and R. Pearson, "Line and Streak Detection on Polished and textured Surfaces using Line Integrals", SPIE/IS&T: Electronic Imaging Symposium, San Francisco, CA, Jan. 2011.
52. M. Jaber, S. Vantaram and E. Saber, "A Probabilistic Framework for Unsupervised Evaluation and Ranking of Image Segmentations", AIPR, Washington, DC, 2010.

53. M. Jaber and E. Saber, "A Bayesian Network-Based Approach for Identifying Regions of Interest Utilizing Global Image Features," Proceedings of the SPIE Optical Engineering + Applications Conference, Vol. 7798, San Diego, CA, Aug. 2010.
54. M. Jaber, E. Saber, M. Shaw, J. Hewitt, "A robust and fast approach for multiple image components stitching," Proceedings of SPIE/IS&T: Electronic Imaging Symposium, San Jose, CA, Jan. 2010.
55. S. R. Vantaram, E. Saber, S. Dianat, M. Shaw and R. Bhaskar, "An Adaptive and Progressive Approach for Efficient Gradient-based Multi-resolution Color Image Segmentation", International Conference on Image Processing, Cairo, Egypt, November 2009.
56. S. Khullar, E. Saber, S. Dianat, J. Trask, R. Lawton, and M. Shaw, "Automatic Multi-resolution Spatio-Frequency Analysis for Print Mottle Evaluation", Proceedings of the SPIE: 17th Color Imaging Conference, Albuquerque, NM, November 2009.
57. P. Gurram, E. Saber, H. Rhody, "Automated 3d Object Identification Using Bayesian Network", AIPR workshop, Washington, D.C., October 2009.
58. X. Fan, H. Rhody and E. Saber, "A Novel Improvement for the HCL-MMI Multi-modal Image Registration by SIFT Algorithm", IEEE Western NY Image Processing Workshop, Sept. 2009.
59. M. Jaber, E. Saber, F. Sahin, "Extraction of Memory Colors Using Bayesian Networks", IEEE SMC International Conference on System of Systems Engineering, Albuquerque, NM, June 2009.
60. S. R. Vantaram, E. Saber, V. Amuso, M. Shaw, R. Bhaskar, "Unsupervised Image Segmentation by Automatic Gradient Thresholding for Dynamic region Growth in the CIE L*a*b* Color Space", SPIE/IS&T: Electronic Imaging Symposium, San Jose, CA, Jan. 2009.
61. X. Fan, H. Rhody and E. Saber, "A Spatial Feature Enhanced MMI Algorithm for Multi-modal Wild Fire Image Registration", AIPR, Washington, DC, 2008.
62. X. Fan, H. Rhody and E. Saber, "A Novel Feature Enhanced MMI based Registration Algorithm for Automated Maps and Images", IEEE International Geoscience & Remote Sensing Symposium, Boston MA, July 2008.
63. L. Garcia, E. Saber, V. Amuso, M. Shaw and R. Bhaskar, "Automatic Color Image Segmentation By Dynamic Region Growth And Multimodal Merging Of Color And Texture Information", International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, March 2008.
64. G. Balasubramanian, E. Saber, V. Mistic, E. Peskin and M. Shaw, "Unsupervised Color Image Segmentation by Dynamic Color Gradient Thresholding" Proceedings of SPIE/IS&T: Electronic Imaging Symposium, San Jose, CA, Jan. 2008.
65. M. Jaber, E. Saber, S. Dianat, and M. Shaw, "Identification and Ranking of Relevant Image Content", Proceedings of SPIE/IS&T: Electronic Imaging Symposium, San Jose, CA, Jan. 2008.
66. P. Gurram, E. Saber, H. Rhody, "Extraction of digital elevation map from parallel-perspective stereo mosaics", Proceedings of SPIE/IS&T: Electronic Imaging Symposium, San Jose, CA, Jan. 2008.
67. X. Fan, H. Rhody and E. Saber, "An Algorithm for Automated Registration of Maps and Images based on Feature Detection and Mutual Information", Proceedings of SPIE/IS&T: Electronic Imaging Symposium, San Jose, CA, Jan. 2008.
68. P. Gurram, E. Saber, H. Rhody, S. Lach and J. Kerekes, "3D Scene Reconstruction through a Fusion of Passive Video and Lidar Imagery", AIPR workshop, Washington, D.C., October 2007.
69. K. Chandu, E. Saber and W. Wu, "A Mutual Information Based Automatic Registration And Analysis Algorithm For Defect Identification In Printed Documents", International Conference on Image Processing, San Antonio, TX, September 2007
70. M. K. Reddy, V. Mistic, E. Saber and J. Trask (HP), "A new Adaptive Edge Enhancement Algorithm for Color Laser Printers", ICASSP, Honolulu, Hawaii, April 2007.
71. X. Fan, H. Rhody and E. Saber, "A Harris Corner Label Enhanced MMI Algorithm for Multi-Modal Airborne Image Registration", International Conf. on Computer Vision Theory and Appl., Barcelona, March 2007.
72. P. Gurram, E. Saber and H. Rhody, "A Novel Triangulation Method for Building Parallel-Perspective Stereo Mosaics", Proceedings of SPIE/IS&T: Electronic Imaging Symposium, San Jose, CA, Jan. 2007.
73. M. Husain, E. Saber, V. Mistic and S. Joralemon, "Dynamic Object Tracking By Partial Shape Matching For Video Surveillance Applications", International Conference on Image Proc., Atlanta, GA, October 2006.
74. X. Fan, H. Rhody and E. Saber, "A Comparison of Exhaustive Search vs. Gradient Search for Automatic Imagery Registration based on MMI", IEEE Western NY Image Processing Workshop, Rochester, NY 2006.
75. O. Ugbeme, E. Saber, and W. Wu, "An Automated Defect Classifying Algorithm for Printed Documents", International Congress on Imaging Science, Rochester, NY, May 2006.

76. H. Santos, E. Saber, and W. Wu, "A New Algorithm for Streak Detection in Mottle and Noisy Images", International Congress on Imaging Science, Rochester, NY, May 2006.
77. V. Sampath, V. Misic, E. Saber, H. Liu, and Y. Yu, "Seed Localization Using Trus And Grf Based Gaussian Filtering For Brachytherapy Applications", International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, 2006.
78. X. Fan, H. Rhody and E. Saber, "Automatic Registration of Multi-Sensor Airborne Imagery", AIPR workshop, Washington, D.C., Oct. 19-21, 2005.
79. K. Modi, F. Sahin and E. Saber, "An Application of Human Robot Interaction: development of a Ping-Pong Playing Robotic Arm", International Conference on SMC, 2005.
80. A. E. Ononye, A. Vodacek and E. Saber, "Extraction of Active Fire Line and Map Using AVIRIS Imagery", EastFIRE Conference, May 11-13, 2005, Fairfax, VA
81. E. Saber, Y. Xu and A. M. Tekalp, "Object Based Image Labeling using Partial Matching Guided Search", International Conf. on Acoustics, Speech, and Signal Proc., Philadelphia, USA, 2005.
82. E. Saber, Y. Xu and A. M. Tekalp, "VOOGLE: Tools for Labeling & Manipulating Partial and Full Objects in Images Invariant to Translations, Rotations, Scale and Reflections", Western New York Workshop, Sept. 2004.
83. Y. Xu, E. Saber, and A. M. Tekalp, "Semantic Object Segmentation by Dynamic Learning from Multiple Examples", International Conf. on Acoustics, Speech, and Signal Proc., May 2004, Montreal, Canada.
84. M. Celik, G. Sharma, A. M. Tekalp, E. Saber, "Lossless Authentication Watermark", SPIE/IS&T: EI 2003, San Jose, CA, USA.
85. M. Celik, G. Sharma, A. M. Tekalp, E. Saber, "Reversible Data Hiding", International Conf. on Image Processing 2002, Rochester, NY, USA.
86. M. Celik, G. Sharma, E. Saber, and A. M. Tekalp, "A Hierarchical Image Authentication Watermark with Improved Localization & Security", International Conf. on Image Processing 2001, Thessaloniki, Greece.
87. M. Celik, E. Saber, G. Sharma, A. M. Tekalp, "Geometry-Invariant Watermarking", SPIE/IS&T: Electronic Imaging, San Jose, 2001.
88. Y. Xu, E. Saber, and A. M. Tekalp, "Contour based Shape matching of Partially Occluded Objects for Image Labeling using Hierarchical Content Description", SPIE/IS&T : Electronic Imaging 2001, San Jose, CA.
89. Y. Xu, E. Saber, and A. M. Tekalp, "Image Retrieval through Shape Matching of Partially Occluded Objects using Hierarchical Content Description", International Conf. on Image Proc., Vancouver, Canada, 2000.
90. P. Duygulu, Y. Xu, E. Saber, A. M. Tekalp, and F. T. Yarman-Vural, "Object Based Image Retrieval for Multi-Level Segmentation", International Conf. on Acoustic, Speech, and Signal Processing, Istanbul, Turkey, 2000.
91. Y. Xu, E. Saber, and A. M. Tekalp, "Object-Based Image Retrieval through Learning from User Search Patterns and Profiles", SPIE/IS&T Electronic Imaging, January 2000, San Jose, California.
92. Y. Xu, E. Saber, and A. M. Tekalp, "Learning-based Hierarchical Content Description for Object Formation and Retrieval", IEEE Image Processing Workshop, Rochester, NY, September 1999.
93. Y. Xu, E. Saber, and A. M. Tekalp, "Object Formation and Retrieval using a Learning-Based Hierarchical Content Description", International Conf. on Image Processing, 1999, Kobe, Japan.
94. Y. Xu, E. Saber, and A. M. Tekalp, "Hierarchical Content Description and Object Formation by Learning", Computer Vision and Pattern Recognition Workshop, Fort Collins, Colorado, 1999.
95. M. Xia, E. Saber, G. Sharma, and A. M. Tekalp, "Total Least Squares Regression in Neugebauer Model Parameter Estimation for Dot-on-Dot Halftone Screens", Non Impact Printing, October 1998.
96. M. Xia, E. Saber, G. Sharma, and A. M. Tekalp, "Total Least Square Technique in Color Printer Characterization", International Conference on Image Processing, Chicago, 1998.
97. M. Xia, E. Saber, G. Sharma, and A. M. Tekalp, "Adaptive Content Dependent Color Rendering of Images and Documents", SPIE/IS&T: Electronic Imaging, San Jose, 1998.
98. M. Xia, E. Saber, G. Sharma, and A. M. Tekalp, "Total least squares regression in color printer calibration", IEEE Image Processing Workshop, Rochester, NY, September 1997.
99. E. Saber and A. M. Tekalp, "Image Annotation and Retrieval by Integrating Color, Shape and Texture", International Conference on Image Processing, September 1996, Lausanne, Switzerland.
100. E. Saber and A. M. Tekalp, "Region-Based Image Annotation using Color and Texture Cues", European Signal Processing Conference, September 1996, Trieste, Italy.
101. E. Saber and A. M. Tekalp, "Detection of Faces and Eyes using Color, Shape and Symmetry-Based Cost Functions", International Conference on Pattern Recognition, August 1996, Vienna, Austria.

102. E. Saber, A. M. Tekalp, and G. Bozdagi, "Fusion of Color and Edge Information for Improved Segmentation and Edge Linking", International Conf. on Acoustics, Speech, and Signal Proc., May 1996, Atlanta, Georgia.
103. E. Saber and A. M. Tekalp, "Image Query-by-Example using Region-Based Affine Shape Matching", SPIE/IS&T Electronic Imaging, Volume 2666, January 1996, San Jose, California.
104. E. Saber, A. M. Tekalp, R. Eschbach and K. Knox, "Annotation Of Natural Scenes Using Adaptive Color Segmentation", SPIE/IS&T Electronic Imaging, February 1995, San Jose, California.

PATENTS, PATENT APPLICATIONS and PATENT PUBLICATIONS

1. D. Cherry, R. Jessome, R. Maggard, D. Siddall, Y. Raja and E. Saber, "Part Replacement Predictions using Convolutional Neural Networks", under review.
2. A. Rangnekar, E. Saber, S. Moudgalya, "Interpolating Pixel Values", Filed February 2016, Patent Pending.
3. P. Lee, E. Saber, O. de Lima, D. Day, P. Bauer, M. Shaw, R. S. Twede, S. Janakiraman, B. Sorensen. "Detecting Document Objects". # PCT/US2015/053762.
4. S. Vantaram, E. Saber, S. Dianat, M. Shaw and R. Bhaskar, "Methods for Adaptive and Progressive Gradient-Based Multi-resolution Color Image Segmentation and Systems Thereof", US 8515171B2.
5. M. Shaw, R. Bhaskar, L. Garcia, E. Saber, V. Amuso, "Image Segmentation using Dynamic Color Gradient Threshold, Texture and Multimodal Merging", US Patent Application Publication US2009/0080773A1.
6. E. Saber, M. Nelson, M. Stauffer, D. Bohan and R. Kothari, "Correcting Artifacts on A Display", US Patent 9,620,082.
7. M. Shaw, R. Bhaskar, G. Balasubramanian, E. Saber, V. Misic, and E. Peskin "Unsupervised Color Image Segmentation by Dynamic Color Gradient Thresholding", US Patent 7,873,214.
8. E. Saber & R. Loce, "Corner Sharpening of Text and Line Art in a Super Resolution Anti-Aliasing Image Path", US Patent 7,536,052.
9. L. K. Mestha, S. Bolte, E. Saber, and S. Updegraff, "Systems and Methods for Obtaining a Spatial Color Profile, and Calibrating a Marking System", US Patent Application Publication 2004/0136013, July 2004.
10. L. K. Mestha, E. Saber, "Systems and methods for sensing marking substrate area coverage using a spectrophotometer", US Patent 7,110,142 and European EP1309176A2.
11. D. Damji, A. Leon, P. Perez and E. Saber, "Processing system for replaceable modules in a digital printing apparatus", European Patent EP1079278B.
12. E. Saber, D. Damji, A. Leon, and P. Perez, "Remanufacturing system for replaceable modules in a digital printing apparatus", US Patent 6173128.

EXTERNAL FUNDING/FUNDED PROPOSALS

Proposals Funded while at RIT

- Acquired/continue to acquire funding with multiple grants from government agencies and corporate partners (Hewlett-Packard, Dataphysics, Lenel, Xerox) in excess of \$3.5 Million as PI and in excess of \$5 Million as PI/Co-PI (total Portfolio) since joining RIT. A complete statement listing the details of each grant can be provided upon request.

Proposals Funded prior to Joining RIT as an employee of Xerox Corporation and University of Rochester:

- A. M. Tekalp and E. Saber (co-PI), "An intelligent visual database system: Hierarchical content description and matching using integrated similarity metrics", funded for \$243K for four years at the University of Rochester by the National Science Foundation under NSF Grant IIS – 9820721.

HONORS AND AWARDS

- Awarded the Prestigious Trustees Scholarship award – the highest award at RIT with regards to research recognition (2012)

- Elected as the Kate Gleason College of Engineering Scholarship award winner (2012).
- Elected as Electrical and Microelectronic Engineering Gleason Professor for 3 years Fall 2011 – Summer 2014.
- Elected as PI Millionaire by RIT Sponsored research organization.
- Winner of an M.S./Ph.D. scholarship for graduate study from Xerox Corporation.
- Winner of the quality recognition award from Xerox Corporation for outstanding performance.
- Core member of a toner/developer facility engineering project team recognized as "The 1991 and 1993 team of the year" by the Delaware Valley Chapter of the Project Management Institute.
- Elected to the Electrical Engineering Honor Society, Eta Kappa Nu.
- Winner of Gibran Khalil Gibran Scholarship for outstanding academic achievements.
- Valedictorian of the Electrical and Computer Engineering Department at the University of Buffalo.
- Recipient of several prizes and awards from the University of Buffalo and the Mohawk Valley Comm. College for excellent academic achievements; and from Xerox Corporation for outstanding performance.

PROFESSIONAL ACTIVITIES

- Senior Member of the Institute of Electrical & Electronic Engineers.
- Member of the IEEE Signal Processing Society.
- Plenary Chair for ICIP 2021.
- Technical Program Chair for ICASSP 2017 in New Orleans.
- General Chair for ICIP 2012.
- Member of the IEEE Signal Processing society conference board for 3 year team starting January 2013.
- Tutorial chair for the International Conferences on Image Processing, ICIP 2007 and ICIP 2009.
- Special session chair on Color Image Processing for the European Signal Processing Conference, Sept. 2006.
- Finance Chair for the International Conference on Image Processing 2002 held in Rochester, NY.
- Former Area Editor for the Journal of Electronic Imaging.
- Guest Editor for the "Color Image Processing" issue of the Signal Processing Magazine.
- Associate Editor for the IEEE Transactions on Image Processing for five years. Term ended: April 2009.
- Former Associate Editor for the IEEE Signal Processing Magazine for DSP Applications Forum.
- Former Member of the IEEE Image & Multidimensional Digital Signal Processing Technical Committee.
- Former Member of IEEE Tech. Comm. on Industry DSP Technology in 2003 & 2004 & Chair for 2005, 2006.
- Technical program committee member for ICIP 2009 and prior ICIPs.
- Technical program committee member for ICASSP 2010 and prior ICASSPs.
- Session chair for several ICIP & ICASSP Conferences and Workshops.
- Chairman, vice-chairman, treasurer, and secretary of the IEEE Rochester Chapter of the signal processing society in 1998, 1997, 1996 and 1995 respectively.
- Reviewer for the IEEE Trans. on Image Processing, IEEE Trans. on Pattern Analysis and Machine Intelligence, Graphical Models and Image Processing, IEEE Trans. on Signal Processing, IEEE Signal Processing Letters, Color Research and Applications, Graphical Modeling and Image Understanding, Image and Vision Computing, Optical Engineering, Journal of Imaging Science and Technology, and the Journal of Electronic Imaging.
- Technical committee for Western New York Imaging Workshop in 1997 and 1999 and general chair in 1998.
- Co-Chair of the Xerox Electronic Image and Video Processing Technology Council.

References Available Upon Request

Dr. Eli Saber, PhD



Mobile: (585) 727-3126

Email: esaber228@gmail.com

Summary of Expert Witness Experience

IPRs/Declarations	Provided declarations and consulting services in support of multiple IPR proceedings (55+) on behalf of Petitioners and Patent Owners.
Depositions	Provided 30 Depositions
Court Testimony	• International Trade Commission Court: Sept. 2024 • International Trade Commission Court: Oct. 2019 • International Trade Commission Court: Sept. 2018
Expert Opinions	Provided multiple expert opinions and reports in support of validity/invalidity, infringement/non-infringement, domestic industry, and rebuttals/reply/sur-reply as needed.
Claim Construction	Provided multiple reports in support of claim construction.
Markman Hearings	Attended one Markman Hearing at the International Trade Commission Court (June 2018) and provided a technology tutorial .

Companies, Law Firms and Third Party Expert Witness Search Firms

Companies that I have been retained on their behalf as an Expert	3Shape, Alarm.com Inc., Amazon, Angel Group Co., Angel Technologies Group LLC, Autonomous Devices, Blackberry, Canon, Corephotonics, Facetec, Genetec, Hewlett-Packard, Hisense, Intellectual Ventures II, Interactive Digital Solutions, Lyft, LNHL, LLC/Seek Now Inc., Malikie Innovations, Maxell, Netflix, Nielsen Company LLC, Nikon, Perfect Corporation, Qomo Hitevision, RealD, Roku, Scandit AG, Sony, Southern Cellular, Truist Bank, Valmarc, VL Collective IP LLC and Wells Fargo.
Law Firms that I have been retained by as an Expert	• Bernstein IP Strategy • Buchanan, Ingersoll and Rooney • Cravath, Swaine & Moore • Davis Polk • Erise Intellectual Property Law • Fabricant, LLP • Fenwick and West • Fish & Richardson • Fitzpatrick, Cella, Harper and Scinto • Hilgers Graben • Holland and Knight, LLP • Lerner, David, Littenberg, Krumholz and Mentlik • Mayer-Brown • McKool-Smith • Mei and Mark • Michelman and Robinson • Mintz, Levin, Cohn, Ferris, Glovsky & Popeo • Morgan, Lewis & Bockius • Morrison & Foerster • Nelson, Mullins, Riley and Scarborough LLP • Norton Rose Fulbright • Oblon, McClelland, Maier and Neustadt • OneLLP • Pepper Hamilton, PC • Perkins-Coie • Quinn Emanuel Urquhart & Sullivan, LLP • Reichman Jorgensen Lehman & Feldberg LLP • Russ, August & Kabat • Skiermont Derby • Steptoe • Troutman Pepper (Formally known as Pepper Hamilton) • Volpe and Koenig PC • Williams & Connolly • Winston & Strawn

Overview of Professional Experience in: 1) Machine Learning/Deep Learning Neural Networks, 2) Digital Video Compression and Coding, 3) Computer Vision and Image/Video Processing, and 4) Industrial Systems (See Curriculum Vitae for details)

Machine Learning/Deep Learning/Artificial Neural Networks

- Experience in machine learning/deep learning techniques and algorithms with applications for Target Detection/Tracking and Activity Recognition from Multimodal data.
- Experience in developing machine learning/deep learning/artificial intelligence type algorithms for fusion of multiple modalities using CNNs to detect and track small/sparse targets in aerial imagery.
- Served as Principle Investigator (PI) on multiple grants from the Department of Defense involving machine learning/deep learning algorithms and techniques.
- Served as an expert on cases involving machine learning algorithms and technologies.
- Developed and currently teaching an advanced graduate course titled: “Advanced Neural Networks and Deep Learning: Theory, Mathematics and Algorithms I”. The course is intended to provide an in-depth comprehensive study of the theory, mathematics and algorithms for advanced neural networks and their usage in deep learning. It is primarily targeted for upper graduate students that are planning to pursue advanced studies/research in the field. The course provides a thorough in-depth study of shallow and deep neural networks architectures and algorithms, activation functions, optimization techniques, initialization strategies, linear and logistic regression, linear two-class and multi-class classification, linear unsupervised learning, feature engineering and selection, fully connected neural networks and backpropagation. The course was taught for the first time in Spring 2024.
- Currently developing a follow-on graduate course to the one outlined above titled: “Advanced Neural Networks and Deep Learning: Theory, Mathematics and Algorithms II”. This course will build on the prior offering and provide in-depth coverage for nonlinear feature engineering, principles of feature learning, kernel methods, tree-based and random forest classifiers, convolutional neural networks, recurrent neural networks, attention mechanisms and transformers, graph neural networks, generative adversarial networks and deep reinforcement learning. This course is intended to debut in Fall 2026.

Digital Video Compression and Coding

- Experience in image and video compression/coding algorithms, techniques and standards including JPEG, MPEG-1, MPEG-2, H.261, H.262, H.263, H.26L, H.264, H.265 and AV-1.
- Served as an expert on multiple IPRs and an ITC case involving video compression/coding technology and standards.
- Testified in ITC court in Sept. 2024 on behalf of Amazon/Hewlett- Packard on technology related to MPEG-1, MPEG-2, H.264, H.265 and AV-1
- Expertise with all technical aspects of video compression techniques such as motion estimation/compensation, segmentation, transformations, quantization, entropy coding, interpolation, color conversion, filtering and optimization.

- Currently working to develop a graduate course titled: “Image and Video Compression”. The course will start by providing a brief history of video coding and video coding standards and then progress to discussing video basics, formats and fundamentals of video compression. It will then progress to cover JPEG, JPEG 2000, H.261, MPEG-1, MPEG-2, H.263, MPEG-4 visual, H.264 and H.265 and if time permits, AV-1 and H.266. The course is intended to debut in Spring 2027.

Computer Vision, Digital Image/Video Processing

- Experience in developing digital image and video processing algorithms and techniques for motion estimation, segmentation, registration, classification, identification, pattern recognition, facial recognition, biomedical image processing, compression, surveillance systems, multiple camera systems, machine learning, deep learning and artificial intelligence for a variety of applications including object recognition and tracking, video analysis, video surveillance, change detection, biomedical, color engineering, document processing, etc.
- Experience in biometrics and face recognition/human segmentation and tracking. Developed/led the development of segmentation algorithms to identify, track and recognize faces/humans in color and multispectral/hyperspectral imagery using single and multiple cameras.
- Experience in computer vision, stereo imaging, 3d scene modeling, object reconstruction from multiple camera imagery, shape recognition/matching, and hierarchical object representations.
- Experience with Lidar, hyperspectral/multispectral imagery and multispectral/multimodal imagery.
- Experience in grayscale and color document segmentation/analysis using cameras, scanners for printing type applications.
- Served as an expert on multiple cases involving computer vision/digital video processing algorithms and technologies.
- Testified in ITC court in Sept. 2018 and Oct. 2029 on behalf of 3Shape on algorithms/technology related to dental imaging.
- Founded and developed the Digital Video Processing (DVP) course at the Electrical Engineering Department at RIT. The DVP course covers pertinent mathematical fundamentals (sampling, structure conversions, etc.), 2d and 3d motion estimation/segmentation, video filtering, stereo imaging, restoration, image/video compression/coding and super resolution.
- Reestablished and Developed the Pattern Recognition course at the Electrical Engineering Department at RIT. The PR course covers pertinent concepts in Bayes decision theory, linear and nonlinear classifiers, artificial neural networks, feature selection/generation/extraction and clustering.
- Established the Image, Video and Computer Vision Laboratory at the Electrical Engineering Department at RIT. Lab has been funded by multiple government grants from the Department of Defense as well as a variety of Industrial grants from companies such as Hewlett-Packard, Dataphysics, Lenel, and Xerox Corporation.

Industrial Experience

- Industrial experience in the design, development, and installation of distributive and programmable logic type control systems hardware/architecture and corresponding software to operate toner production facilities that included batch feeders, extruders, grinders, classifiers, screeners, and packaging systems with multiple instruments such as light sensors, level problems, proximity switches, temperature sensors and other types of instrumentation. Worked on a team responsible for building two large toner production facilities that manufactured ~1,000 lbs of toner per hour.
- Industrial experience in designing and developing architectures/imaging pipelines for digital front ends intended to drive high speed/high quality full process color print engines that involve scanners, LED bars, camera systems and electrophotographic technologies.
- See Curriculum Vitae for additional details.

Active Cases

[C.17] Advanced Coding Technologies, LLC vs Apple and Google

March 2025 – Present

- Case No. 2:24-cv-00353-JRG
- Legal Firm: Fabricant, LLP
- Retained as an expert by the Barr Group on behalf of Advanced Coding Technologies, LLC to provide expert opinions with regards to claim construction, infringement and invalidity.
- [Provided a Deposition on July 16, 2025.](#)
- [Provided a Deposition on September 15, 2025.](#)

[IPR.32] Snap v. Nokia Corp.

May 2025 – Present

- Legal Firm: Winston & Strawn LLP
- Retained as an expert on behalf of Snap to provide assessments and expert opinions on IPRs
- Provided expert opinions and a declaration in support of IPR-2025-01114 (US 8,175,148)
- Provided expert opinions and a declaration in support of IPR-2025-01338 (US 7,724,818)
- Provided expert opinions and a declaration in support of IPR-2025-01345 (US 9,036,701)

[IPR.31] Perfect Corporation v. Zugara LLC

April 2025 – Present

US Patent and Trademark Office, Patent Trial and Appeal Board

- Legal Firm: Bernstein IP Strategy
- Retained as an expert by Bernstein IP Strategy on behalf of Perfect Corporation.
- Provided expert opinions and a declaration in support of IPR-2025-01144 (US 10,482,517)

[IPR.30] Element TV Co. v. Nokia Corp.

May 2025 – Present

- Legal Firm: Perkins-Coie
- Retained as an expert on behalf of Element TV Co to provide assessments and expert opinions on Potential IPRs
- Provided expert opinions and a declaration in support of IPR-2025-01068 (US 8,050,321)
- Provided expert opinions and a declaration in support of IPR-2025-01109 (US 7,532,808)

[IPR. 29-2X] Tesla Inc. v. Intellectual Ventures II LLC

March 2025 – Present

- Legal Firm: Volpe and Koenig PC
- Retained as an expert by IMS Consulting and Expert Services on behalf of Intellectual Ventures II LLC to provide expert opinions as needed
- Provided expert opinions and a declaration in support of IPR-2025-00222 (US 9,232,158)

- Provided expert opinions and a declaration in support of IPR-2025-00339 (US 7,916,180)

[C.16/IPR.28 – 6X] Autonomous Devices LLC v. Tesla, Inc

Oct. 2022 - Present

US District Court, District of Delaware, case No.

- Legal Firm: McKool-Smith
- Retained as an expert by the Barr Group on behalf of Autonomous Devices.
- Provided declarations including Motion to Amend on multiple IPRs (6X): IPR2023-01054, IPR2023-01055, IPR2023-01169, IPR2023-01172, IPR2023-01173 and IPR2024-00745 and Ex-Parte reviews
- Provided a Deposition on May 16th, 2024 with regards to IPR 2023-01054.
- Provided a Deposition on May 21th, 2024 with regards to IPR 2023-01055.
- Provided a Deposition on June 14th, 2024 with regards to IPR 2023-01169.
- Provided a Deposition on June 21st, 2024 with regards to IPR2023-01172 and IPR 2023-01173.
- Provided a Deposition on April 4th, 2025 with regards to IPR2024-00745
- Provided a Deposition on June 27, 2025 with regards to IPR2024-00745

[C.15/IPR.27] RealD, LLC v. Microsoft Corp., NO. 2:22-cv-00942-TL

August 2025 – Present

- Legal Firm: OneLLP
- Retained as an expert on behalf of RealD, LLC
-
- Legal Firm: Michelman & Robinson
- Retained as an expert on behalf of RealD, LLC
- Provided expert opinions and a declaration in support of IPR2023-01046. Patent Owner Preliminary Response filed on 10/13/2023.
- Provide expert opinions and additional reports in support of Federal case.
- Provide algorithm support for software review.
- Provided expert opinions and a declaration in support of Patent Owner Response with regards to IPR2023-01046. POR filed on March 15th, 2024
- Provided a Deposition on June 5th and 7th, 2024 with regards to IPR 2023-01046.

Feb . 2023 – June 2024

Inactive Cases

[C.14] In Re Lyft Rideshare Litigation

Dec. 2021 – Present

Superior Court of the State of California, Case No. CJC-20-005061

- Legal Firm: Williams & Connolly, LLP
- Retained as an expert by WIT Legal, LLC on behalf of Lyft (Defendant)
- Provide reports and consulting services on the challenges of the use of Dashcam solutions in Vehicles
- [Provided a Deposition on March 24th, 2023.](#)
- The portion of the Case that requires my expertise has been settled.

Concluded Cases

[IPR.26] Meta Platforms, Inc.; Instagram, Inc.; WhatsApp LLC; Meta Platforms Technologies, LLC; and Giphy, Inc. (Petitioners) v. VL Collective IP LLC (Patent Owner)

Jan. 2024 – Present

- Legal Firm: Reichman, Jorgensen, Lehman & Feldberg LLP
- Retained as an expert on behalf of VL Collective IP LLC
- Provided expert opinions, a declaration and a supplemental declaration in support of Patent Owner Response with regards to IPR2023-00924.
- [Provided a Deposition on May 30th, 2024 with regards to IPR 2023-00924.](#)
- [Provided a Deposition on August 14, 2024 with regards to IPR 2023-00924.](#)

[C.13] Malikie Innovations Limited and Key Patent Innovation Limited v. Aser & Asus

Sept. 2024 – April 2025

- Legal Firm: Reichman, Jorgensen, Lehman and Feldberg LLP
- Retained as an expert on behalf of Malikie Innovations Limited and Key Patent Innovations Limited
- Provide expert opinions in support of legal proceedings.
- Case was dismissed in April 2025

[C.12/IPR.25-3X] Amazon v. Nokia

Jan. 2024 – April 2025

- Legal Firm: Perkins-Coie, Morgan Lewis and Fenwick and West
- IPR Proceedings:
 - Retained as an expert on behalf of Amazon (Petitioner)
 - Provided expert opinions and a declaration for an IPR proceeding re. US 7,724,818 (IPR2024-00572)
 - [Provided a Deposition on January 24, 2025 for IPR2024-00572.](#)
 - Provided expert opinions and a declaration for an IPR proceeding re. US 6,856,701
 - Provided expert opinions and a declaration for an IPR proceeding re. US 8,175,148
 - Case was settled in April 2025. All IPRs have been terminated.
- ITC Case No. 337-TA-1380:
 - Provided expert reports with regards to invalidity, non infringement and domestic industry with regards to ITC case No. 337-TA-1380
 - [Provided a Deposition on June 28st, 2024 for ITC case No. 337-TA-1380.](#)
 - [Provided ITC trial testimony on September 10-11th, 2024 \(Trial was held Sept. 09-13, 2024\) with regards to Invalidity, Non-infringement and Domestic Industry.](#)

[IPR.24] Hand Held Products inc. (Plaintiff) v. Scandit AG (Defendant)

Nov. 2024 – March 2025

- Legal Firm: Quinn, Emanuel, Urquhart & Sullivan, LLP
- Retained as an expert on behalf of Scandit AG to provide expert opinions in a potential Inter Partes Proceedings regarding US 11,809,952, US 10,803,274 and US 11,347,958
- Case was settled in March 2025

[IPR.23] Dental Imaging Technologies Corporation (Petitioner) v. 3Shape A/S (Patent Owner)

August. 2024 – January 2025

US Patent and Trademark Office, Patent Trial and Appeal Board

- Legal Firm: Buchanan, Ingersoll and Rooney PC
- Retained as an expert on behalf of 3Shape A/S to provide expert opinions in an inter partes proceeding involving U.S. Patent US 10,695,151 (IPR2024-00849).
- Provided a declaration in support of Patent Owner.
- Case has been settled

[IPR.22-6X] 3Shape A/S (Petitioner) v. Medit Corporation

Jan. 2024 – Jan. 2025

US Patent and Trademark Office, Patent Trial and Appeal Board

- Legal Firm: Buchanan, Ingersoll and Rooney PC
- Retained as an expert on behalf of 3Shape A/S to provide expert opinions and declarations for multiple inter partes proceedings involving U.S. Patent US 10,076,391 (IPR2024-00548), US 11,234,798 (IPR2024-00549), US 9,191,648 (IPR2024-01268), US 9,262,864 (IPR2024-01265 & IPR2024-01266), US 7,912,257 (IPR2024-01260) and US 9,245,374 (IPR2024-01276).
- Cases have been settled

[IPR.21-3X] 3Shape A/S (Petitioner) v. Dental Imaging Technologies & Carestream Dental

Jan. 2024 – January 2025

US Patent and Trademark Office, Patent Trial and Appeal Board

- Legal Firm: Buchanan, Ingersoll and Rooney PC
- Retained as an expert on behalf of 3Shape A/S to provide expert opinions on inter partes proceedings involving U.S. Patent US 10,076,391 (IPR2024-00548), US 11,234,798 (IPR2024-00549) and US 9,191,648
- Cases have been settled

[IPR.20-3X] Medit USA, Crown Jade Co. Ltd. (Petitioner) v. 3Shape A/S (Patent Owner)

Oct. 2023 – January 2025

US Patent and Trademark Office, Patent Trial and Appeal Board

- Legal Firm: Buchanan, Ingersoll and Rooney PC
- Retained as an expert on behalf of 3Shape A/S to provide expert opinions on IPR proceedings involving U.S. Patent Nos. 10,695,151; 11,076,146; 11,368,667.
- Provide expert opinions in support of Patent Owner Response.

- Institution was denied for 11,076,146; 11,368,667.
- Institution granted for U.S. 10,695,151 (IPR2023-01315)
- Provided a declaration in support of Patent Owner regarding US 10,695,151 (IPR2023-01315)
- [Provided a Deposition on August 8th, 2024 for IPR2023-01315.](#)
- Case has been settled

[IPR.19] TVISION INSIGHTS v. The NIELSEN Company (US), LLC

Aug. 2023 – Aug. 2024

US Patent and Trademark Office, Patent Trial and Appeal Board

- Legal Firm: Erise Intellectual Property Law
- Retained as an expert on behalf of the Nielsen Company (US), LLC
- Provided expert opinions and a declaration in support of patent owner in IPR2023-01014. Documents were filed on 10/11/2023.
- PTAB decision to institute was reached January 2024.
- [Provided a Deposition on July 2nd, 2024.](#)

[C.11] Maxell, LTD. v. TCL Electronics Holdings Ltd. Et. al.

June 2024 – Sept. 2024

Case 5:23-CV-00108-RSW-JBB (EDTX) and related proceedings including against LG Electronics and Samsung

- Legal Firm: Mayer-Brown
- Retained as an expert on behalf of Maxell (Plaintiff)
- Provide expert opinions in support of legal proceedings.
- Provided an expert report in support of Plaintiff claim construction
- [Provided a Deposition on August 20th, 2024 with regards to claim construction.](#)

[IPR.18 – 4X] Meta Platforms, Inc. f/k/a Facebook, Inc. v. Angel Technologies Group LLC

May 2023 – October 2023

US District Court, District of Nevada

- Legal Firm: McKool-Smith
- Retained as an expert by Alliance Technical Staffing on behalf of Angel Technologies Group LLC.
- Provide expert opinions and declarations in support of four IPR proceedings (IPR2023-00057, IPR2023-00058, IPR2023-00059, IPR2023-00060).
- [Provided a Deposition on October 20th, 2023 for IPR2023-00057, IPR2023-00058 and IPR2023-00059](#)
- [Provided a Deposition on October 23rd, 2023 for IPR2023-00060.](#)

[C.10] Alarm.com Inc. v. Vivint, Inc NO. 2:23-CV-00004-JRG-RSP

Mar. 2023 – Dec. 2023

- Legal Firm: Cravath, Swaine and Moore, LLP
- Retained as an expert on behalf of Alarm.com, Inc.
- Case has been settled

[C.9] HL Holdings, Inc. v. LNHL, LLC, et al.

July 2023 – Dec. 2023

- Legal Firm: Nelson, Mullins, Riley and Scarborough LLP
- Retained as an expert by IMS Consulting and Expert Services on behalf of LNHL, LLC and Seek Now Inc. f/k/a Ladder Now LLC.
- Provided expert opinions and a report in support of legal proceedings.
- [Provided a Deposition on November 29th, 2023](#)
- Case has been settled.

[IPR.17] Angel Group Co. Ltd. V. Sensen Networks Group Pty Ltd

July 2023 – Oct. 2023

- Legal Firm: Norton Rose Fulbright
- Retained as an expert on behalf of Angel Group Co. Ltd.
- Provide expert opinions in support of a potential IPR proceeding.
- Case has been settled.

[C.8] United Services Automobile Association v. Truist Bank

Mar. 2023 – Oct. 2023

- Case No. 2-22-CV-00291
- Legal Firm: Cravath, Swaine and Moore, LLP
- Retained as an expert on behalf of Truist Bank.
- Performed review of six patents.
- Case has been settled.

[IPR.16 – 3X] Dental Imaging Technologies Corp. (Petitioner) v. 3Shape A/S (Patent Owner)

Feb. 2023 – July 2023

US Patent and Trademark Office, Patent Trial and Appeal Board

- Legal Firm: Buchanan, Ingersoll and Rooney PC
- Retained as an expert on behalf of 3Shape A/S.
- Provided consulting services in support of POPR for IPR2023-00387 (US 11,076,146).
- Provided consulting services in support of POPR for IPR2023-00400 (US 10,695,151).
- Provided consulting services in support of POPR for IPR2023-00411 (US 11,368,667).

[IPR.15 – 2X] Nike v. Valmarc

January 2023 – Present

- Legal Firm: Hilgers Graben
- Retained as an expert on behalf of Valmarc Corporation
- Provided expert opinions in support of Patent Owner Preliminary Response on petitions IPR2022-01403 & IPR2022-01404
- PTAB granted institution but client opted not to pursue defending the patents.

[C.7] Sensormatic Electronics, LLC v. Genetec, Inc. et al.

Oct. 2021 – Feb. 2023

US District Court, District of Delaware, Case 1:20-cv-00760-MN

- Legal Firm: Fish & Richardson, P.C.

- Retained as an expert by IMS Expert Services on behalf of Genetec, Inc. et al. (Defendant)
- Representing Genetec (USA) Inc. and Genetec Inc. (C. A. No. 20-760)
- Provided Opening, Rebuttal and Reply reports regarding invalidity & Non-infringement with regards to US 7,307,652.
- [Sat in on a deposition of Plaintiff expert \(April 29th, 2022\) per counsel request.](#)
- [Provided a deposition on May 3rd, 2022](#)

[IPR.14-2X] Roku, Inc. v. Flexiworld Technologies, Inc.

June 2022 –

October 2022

US Patent and Trademark Office, Patent Trial and Appeal Board, Case No. IPR2022-01112

US Patent and Trademark Office, Patent Trial and Appeal Board, Case No. IPR2022-001269

- Legal Firm: Perkins-Coie, LLP
- Retained as an expert by WIT Legal, LLC on behalf of Roku (Petitioner).
- Provided a declaration in June 2022 in support of the IPR petition for US 10,140,073.
- Provided a declaration in July 2022 in support of the IPR petition for US 7,944,577.
- Case was settled in October 2022.

[O.3] Fish & Richardson, P.C.

May 2022 – June 2022

- Legal Firm: Fish & Richardson, P.C.
- Retained as an expert by IMS Expert Services
- Provided expert opinions with regards to a pre-IPR petition

[IPR.13] CareView Communications v. Interactive Digital Solutions, Inc.

Jan. 2022 – April 2022

- Legal Firm: Mintz, Levin, Cohn, Ferris, Glovsky & Popeo, P. C.
- Retained as an expert by IMS Expert Services on behalf of Interactive Digital Solutions, LLC (Defendant) to provide declarations in support of 3 IPR proceedings
- Case was settled in April 2022.

[IPR.12] Sovereign Peak Ventures, LLC v. Hisense Co., LTD

Nov. 2021 – Dec. 2021

US District Court, Eastern District of Texas, Case No. 21-cv-00209

- Legal Firm: Perkins-Coie, LLP
- Retained as an expert by WIT Legal, LLC on behalf of Hisense Co. (Defendant).
- Working to provide a declaration in support of the IPR petition for US 8,019,169.
- Case was settled in Early 2022.

[IPR.11] Perfect Corporation v. Lennon Image Technologies, LLC

Sept. 2021 – June 2022

US Patent and Trademark Office, Patent Trial and Appeal Board, IPR2021-00124

- Legal Firm: Perkins-Coie, LLP
- Retained as an expert by WIT Legal, LLC on behalf of Perfect Corporation (Petitioner).
- Provided a declaration in support of the IPR petition for US 6,624,843.
- Petitioner filed a motion to dismiss its petition on April 5th, 2022.

[IPR.10 – 2X] Apple Inc. v. Corephotonics, LTD

Jan. 2021 – May 2021

- Legal Firm: Russ, August & Kabat,
- Retained as an expert by WIT Legal, LLC on behalf of Corephotonics in two IPR cases.
- Provided two declarations in support of IPR Petitions IPR2020-00487 & IPR2020-00860.
- [Provided a deposition on 04/14/2021.](#)

[O.2] Davis Polk

Aug. 2020 – Nov. 2021

- Legal Firm: Davis Polk
- Retained as an expert by WIT Legal, LLC
- Provided expert and consulting services for video and point cloud type imagery.

[O.1] Cravath, Swaine & Moore, LLP

Feb. 2020 – Mar. 2020 and Jan. 2021

- Legal Firm: Cravath, Swaine & Moore, LLP
- Retained as an expert by WIT Legal, LLC
- Evaluated computer vision technology for various clients.
- No reports, depositions or expert testimonies were provided.

[IPR.9 – 2X] Wells Fargo & Co. v. Lighthouse Consulting Group

Feb. 2020 – Sept. 2020

US Patent and Trademark Office, Patent Trial and Appeal Board, IPR2020-00758 and IPR2020-00759

- Legal Firm: Winston & Strawn LLP
- Retained as an expert by DOAR on behalf of Petitioner Wells Fargo.
- Submitted two declarations in support of IPR Petitions IPR2020-00758 & IPR2020-00759 for US 8,590,940 and US RE44,274.

[C.6] 3Shape A/S v. Align Technology, Inc.

Jan. 2020 – April 2020

US District Court, District of Delaware, Case No. 18-cv-00886 (consolidated with 18-cv-00697)

- Legal Firm: Buchanan, Ingersoll and Rooney, PC
- Retained as an expert by DOAR on behalf of plaintiff for US Patents-in-Suit Nos. 9,962,244 and 9,629,551 with regards to invalidity.
- Case was transferred to Troutman Pepper LLP very shortly after my engagement.

[IPR.8 – 2X] 3Shape A/S et al v. Align Technology, Inc.

Oct. 2019 – Oct. 2020

US Patent and Trademark Office, Patent Trial and Appeal Board, IPR2020-00222 and IPR2020-00223

- Legal Firm: Buchanan, Ingersoll and Rooney, PC
- Retained as an expert by DOAR on behalf of Petitioner 3Shape A/S with respect to US 7,156,661
- Provided two declarations in support of IPR petitions (IPR2020-00223 and IPR2020-0222) for US 7,156,661. Both petitions were filed in December 2019.
- [Provided a deposition on July 30th, 2020 in support of IPR2020-00223 that was instituted by PTAB and supporting materials for follow on responses.](#)

[IPR.7] Uniloc 2017 LLC v. Netflix, Inc. (Petitioner) (18-cv-2055)

July 2019 – July 2020

US Patent and Trademark Office, Patent Trial and Appeal Board, IPR2020-00044

- Legal Firm: Perkins-Coie, LLP
- Retained as an expert by DOAR on behalf of Petitioner Netflix, Inc.
- Provided an expert declaration in support of IPR2020-00044 for US 6,584,229.
- Case was settled in July 2020.

[C.5] Align Technology, Inc. v. 3Shape A/S et al

Feb. 2019 – Feb. 2022

US District Court, District of Delaware, Case No. 17-cv-01646 (LPS) (CJB)

- Legal Firm: Troutman Pepper LLP
- Retained as an expert by WIT Legal, LLC on behalf of Defendant 3Shape for US Patent-in-Suit No. 9,510,757 ('757 Patent).
- Served as expert for invalidity and non-Infringement for US 9,510,757
- Provided expert reports regarding Invalidity and non-infringement of '757
- [Provided a deposition on 10 January 2020](#)
- Case was settled in February 2022.

[C.4] Certain Dental and Orthodontic Scanners and Software

Feb. 2019 – Oct. 2019

US International Trade Commission, Investigation No. 337-TA-1144 (3Shape v. Align)

- Legal Firm: Pepper Hamilton, PC
- Retained as an expert by DOAR on behalf of respondent 3Shape A/S with respect to US Patent-in-Suit No. 9,299,192 ('192 Patent).
- Served as expert on invalidity, non-infringement and domestic industry.
- Provided expert opinions and rebuttal reports in support of the above proceedings.
- [Provided a deposition \(8/13/2019\).](#)
- [Provided trial testimony on October 30th, 2019 \(Trial was held Oct. 24-30, 2019\) with regards to Invalidity, Non-infringement and Domestic Industry.](#)

[IPR.6 - 3X] 3Shape A/S et al v. Align Technology, Inc.

Oct. 2018 – Sept. 2019

US Patent and Trademark Office, Patent Trial and Appeal Board, IPR2019-00132 and IPR2019-00134 and IPR2019-00148

- Legal Firm: Buchanan, Ingersoll and Rooney, PC
- Retained as an expert by DOAR on behalf of respondent 3Shape A/S with respect to US 7,112,065, US 9,451,873 and US 9,299,192
- Represented 3Shape (Petitioner) in three IPR Petitions v. Align (Patent Owner).
- Provided three declarations in support of the IPR petitions for US 7,112,065, US 9,299,192 and US 9,451,873. All three petitions were filed in November 2018 and all three were instituted.
- [Provided three depositions with regards to US 7,112,065 \(8/16/2019\), US 9,451,873 \(09/18/2019\) and US 9,299,192 \(09/30/2019\).](#)

[IPR.5 - 3X] Blackberry Corp. et al v. Maxell, Ltd

Sept. 2018 – Oct. 2018

US Patent and Trademark Office, Patent Trial and Appeal Board, IPR2019-00087 and IPR2019-00095 and IPR2019-00097

- Legal Firm: Oblon, McClelland, Maier and Neustadt, LLP
- Retained as an expert on behalf of petitioner Blackberry Corp.
- Provided three declarations in support of the IPR petitions for US 9,100, 604 (Joinder), US 6,765, 616 and US 7,995,897.
- Case settled in May 2019.

[IPR.4] Exocad GmbH et al v. 3Shape A/S

March 2018 – Sept. 2019

US Patent and Trademark Office, Patent Trial and Appeal Board, IPR2018-00788

- Legal Firm: Buchanan, Ingersoll and Rooney, PC
- Retained as an expert by DOAR on behalf of 3Shape A/S in defending an IPR filings of EXOCAD (Petitioner) v. 3Shape (Patent Owner). Case No. IPR 2018-00788
- Provided a declaration in support of the Patent Owner response for US 9,336,336.
- [Provided a deposition on April 1st, 2019.](#)

[C.3] Certain Intraoral Scanners and Related Hardware and Software

Feb. 2018 – Sept. 2018

US International Trade Commission, Investigation No. 337-TA-1090 & No. 337-TA-1091 (Align v. 3Shape)

- Legal Firm: Pepper Hamilton, LLP
- Retained as an expert by DOAR on behalf of respondent 3Shape A/S.
- Provided expert opinions regarding claim construction, noninfringement, and the domestic industry analysis.
- Attended a *Markman* hearing and provided a tutorial for the technology in suit.
- [Provided a Deposition on July 16, 2018.](#)
- [Provided trial testimony as an expert on Sept. 20th, 2018 with regards to Non-infringement and Domestic Industry.](#)

[IPR.3] Nikon Corporation et al v. ASML Netherlands BV et al

July 2017 –

Nov. 2017

US Patent and Trademark Office, Patent Trial and Appeal Board, IPR2018-00227

- Legal Firm: Morrison & Foerster
- Retained as an expert on behalf of petitioner Nikon Corporation.
- Provided an expert declaration in support of IPR petition for US 7,209,167.

[C.2] Raster Printing Innovations, LLC v. Hewlett-Packard Company et al

April 2016 – May 2016

US District Court, Northern District of Illinois, Case No. 15-cv-03967

- Legal Firm: Morgan, Lewis & Bockius, LLP
- Retained as an expert by IMS Expert Services on behalf of defendant Hewlett Packard.
- Contributed to the development of initial invalidity contentions charts.
- Case settled very quickly after my engagement in June 2016.

[IPR.2] Qomo Hitevision LLC v. Pathway Innovations and Technologies, Inc. Nov. 2015 – Jan. 2016

US Patent and Trademark Office, Patent Trial and Appeal Board, IPR2016-00661

- Legal Firm: Mei & Mark, LLP
- Retained as an expert on behalf of petitioner Qomo Hitevision.
- Provided an expert declaration in support of petition for inter partes review.

[IPR.1] Sony Corp of America. v. Image Processing Technologies, LLC Aug. 2014 – Nov. 2014

US Patent and Trademark Office, Patent Trial and Appeal Board

- Legal Firm: Lerner, David, Littenberg, Krumholz & Mentlik LLP
- Retained as an expert on behalf of Sony.
- Provided an expert declaration in support of petition to institute inter partes review of US Patent No. 6,959,293.

[C.1] Image Processing Technologies, LLC v. Canon Inc. et al June 2011 – Dec. 2012

Case: US District Court, Eastern District of New York, Case No. 10-cv-03867

- Legal Firm: Fitzpatrick, Cella, Harper & Scinto
- Retained as an expert on behalf of defendant Canon Inc.
- Traveled to Japan twice to meet with Canon Technical & Engineering Teams
- Provided reports related to Invalidity, Non-Infringement, Rebuttals, and summary judgement.
- [Provided a Deposition in April 2012.](#)

Industrial Training

Xerox Corporation

Summers of 2006, 2007 and 2008

- Developed and conducted industrial training at Xerox Corporation for engineering personnel over three separate summer periods in the areas of signal and image processing, digital printing, color engineering, device characterization/calibration, rendering, control systems and electrophotography with a distinct focus on digital front ends and print engines.
- Created and taught four distinct modules as follows:
 - Module 1: Fundamentals of Multi-dimensional Signal Processing, Matrix Fundamentals and MATLAB for Digital Printing and Electrophotography Applications. This is a pre-requisite for Modules 2, 3 and 4.
 - Module 2: Color Engineering & Device Characterization/Calibration Techniques for Digital Printing and Electrophotography Applications.
 - Module 3: Image Processing and Rendering Techniques for Digital Printing and Electrophotography Applications.
 - Module 4: Modern Control Techniques for Digital Printing and Electrophotography Applications.

Leica Corporation

May 2006

- Provided pattern recognition/shape matching consulting expertise for Leica corporation.