18.7 A Multimedia Semantic Analysis SoC (SASoC) with Machine-Learning Engine

Tse-Wei Chen, Yi-Ling Chen, Teng-Yuan Cheng, Chi-Sun Tang, Pei-Kuei Tsung, Tzu-Der Chuang, Liang-Gee Chen, Shao-Yi Chien

National Taiwan University, Taipei, Taiwan

Advances in semiconductors and developments in machine learning [1] have led to versatile multimedia applications with semantic processing abilities. Real-time applications, such as face detection, facial-expression recognition, scene analysis [2] and object recognition [3], have become indispensable functionality for Consumer Electronic (CE) products. To deal with complicated video-processing algorithms for multimedia content analysis, many powerful processors have been reported [2-5]. Although these processors can speed up video-processing tasks with massively parallel processing elements, they only focus on the feature-extraction parts, and there is no specialized hardware to support different kinds of advanced machine-learning algorithms, which require extensive computations. In this paper, a Semantic Analysis SoC (SASoC) that accelerates video processing and machine learning simultaneously, is developed to meet the demands of the near future.

The SASoC is characterized as follows. (1) It integrates an Image-Stream Processing System (ISPS) supporting pixel-level feature-extraction operations and a Feature-Stream Processing System (FSPS) supporting vector-level machine-learning algorithms for versatile semantic-analysis applications. (2) Hierarchical memory organization and stream network design make the 2 high-parallelism processing units of ISPS work in a pipeline manner with high hardware utilization. (3) The FSPS can support advanced machine-learning algorithms with high throughput by use of a hierarchical 3-level stream vector-processor architecture. (4) A dynamic frequency scaling technique for multiple clock domains reduces power consumption by 65% by dynamically balancing the loading. (5) Implementation results show that the SASoC provides high performance and a high power efficiency of 671GOPS/W, which outperforms previous systems.

Figure 18.7.1 shows the SASoC architecture, which contains 3 clock domains. The System Monitor adopts the power-aware frequency scaling technique to balance the computational time between ISPS and FSPS and reduces the power consumption. The clock speed of ISPS and FSPS can be adjusted dynamically to satisfy the different requirements of multimedia applications. The ISPS consists of a complete system platform for parallel image processing, and the extracted image features can be sent to FSPS for semantic analysis. As a Machine-Learning Engine, the FSPS contains a 3-level Vector Processing Unit (VPU) to handle high-dimensional feature vectors for different machine-learning algorithms, such as: AdaBoost, Artificial Neural Network (ANN), Support Vector Machine (SVM) and Gaussian Mixture Model (GMM).

Figure 18.7.2 shows the ISPS architecture, which includes a system platform with Sequencer, Slice Memory and Reconfigurable Image Stream Processor (RISP). The Sequencer manipulates the data transmission between the Slice Memory and RISP. After receiving the instructions from the Sequencer, the Slice Memory sends 128b pixel data streams to RISP for video processing. The image data are arranged and stored in 16 banks of Slice Memory, which can continuously provide 16-pixel stripes with arbitrary positions. The RISP, which can process 16×16 window-based operations in 1 cycle, has 4 configuration modes with the 2 processing units, Linear Processing Unit (LPU) and Order Processing Unit (OPU). Both LPU and OPU have Local Pixel Memory to provide 102.4GB/s bandwidth in total, and the processed images and features can be stored in the dual Output Memory of RISP. As shown in Figure 18.7.2, with the Stream Network, LPU and OPU can simultaneously perform in a pipeline manner in Mode C and Mode D, where high hardware utilization can be achieved.

Figure 18.7.3 shows the FSPS architecture, which is a Machine-Learning Engine that contains a Vector Processing Unit (VPU) and a K-Nearest Neighbor (K-NN) Processor. The VPU has a 3-level hierarchical architecture that can process 256 dimensions of vectors in parallel, and operations such as vector inner product, vector distance and exponential computation can be executed in 1 cycle. Each level of VPU has a Local Vector Memory (LVM) for rapid data access and supporting different operations and parallelism. The LVM of the Low-Level VPU and Input Vector Memory (IVM) provide 76.8GB/s bandwidth to Vector ALUs, and input vectors can be sent to different levels of the VPU according to application requirements. Connected to High-Level VPU, the K-NN Processor is designed for the computation of rankings of vector distances, and 128 PEs can sort and store the distances in the same clock cycle.

Example applications based on the SASoC are illustrated in Figure 18.7.4. The first application is concept-based image retrieval, which adopts the concept categories to perform semantic analysis in images, and the real-time retrieval results can be used for scene recognition and photo classification in CE products. The color and texture features are extracted by OPU and LPU, respectively, and GMM-based classification can be accomplished using 3 levels of VPU. Finally, the K-NN Processor computes the nearest neighbor of the captured image and gives retrieval results with the frame rate of 156fps in 160×120 resolution. The second application is face detection, which is widely applied in DSCs and camcorders. After noise reduction from OPU, the Haar-like features are extracted by LPU and sent to the FSPS for classification. The 2 levels of VPU are used to execute the AdaBoost algorithm, and the results of face detection are stored in Output Vector Memory (OVM) with the frame rate of 294fps in 160×120 resolution.

The performance analysis with different single-test operations of the ISPS and FSPS is shown in Figure 18.7.5. In the ISPS, the maximum input data rate is 76.8Gpixel/s when OPU and LPU work in pipeline, and the frame rate is 17,500× higher than the state-of-the-art PC when the frequency of the ISPS is more than 10× slower than a Pentium CPU. In the FSPS, the SVM classification operation reaches 51.2Gdimension/s, which is 164× faster than the PC. The input data rate of database in K-NN operation, including distance calculation, is adaptive to the vector dimension, and the maximum speed is 0.2Gvector/s, which is 11,800× faster than the PC.

In most applications, the computational time for video processing and machine-learning algorithms is different, and the bubble cycles result in redundant power consumption. The comparison of the power-aware frequency scaling technique, which dynamically scales the frequencies of the 2 systems, is shown in Figure 18.7.6. By decreasing the frequency of the FSPS, power consumption can be reduced by 65% without scaling the supply voltage, and the clock signal of either the FSPS or ISPS can be gated if only one system is active.

Figure 18.7.6 also shows the summary of chip features and the comparison with related works [2-5]. The SASoC is fabricated in 90nm CMOS and occupies 28mm^2 with 3M gates and 149KB on-chip SRAM. The die micrograph is shown in Figure 18.7.7.

Acknowledgements:
We thank TSMC University Shuttle Program and Morly Hsieh for process support. We also thank Chip Implementation Center (CIC) for design flow supporting and chip testing. This work is funded by National Science Council and TSMC.

References:
[1] Ethem Alpaydin, *Introduction to Machine Learning*, MIT Press, 2004.
[2] A. Abbo, et al., "XETAL-II: A 107 GOPS, 600mW Massively-Parallel Processor for Video Scene Analysis," *ISSCC Dig. Tech. Papers*, pp. 270-271, Feb. 2007.
[3] Kwanho Kim, et al., "A 125GOPS 583mW Network-on-Chip Based Parallel Processor with Bio-inspired Visual Attention Engine," *ISSCC Dig. Tech. Papers*, pp. 308-309, Feb. 2008.
[4] Sumito Arakawa, et al., "A 512GOPS Fully-Programmable Digital Image Processor with full HD 1080p Processing Capabilities," *ISSCC Dig. Tech. Papers*, pp. 312-313, Feb. 2008.
[5] Joo-Young Kim, et al., "A 201.4GOPS 496mW Real-Time Multi-Object Recognition Processor with Bio-Inspired Neural Perception Engine," *ISSCC Dig. Tech. Paper*, pp. 150-151, Feb. 2009.

978-1-4244-6034-2/10/$26.00 ©2010 IEEE

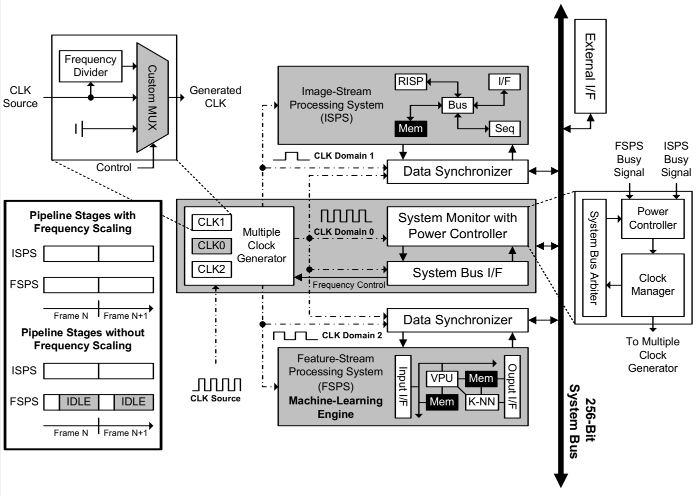


Figure 18.7.1: SASoC Architecture.

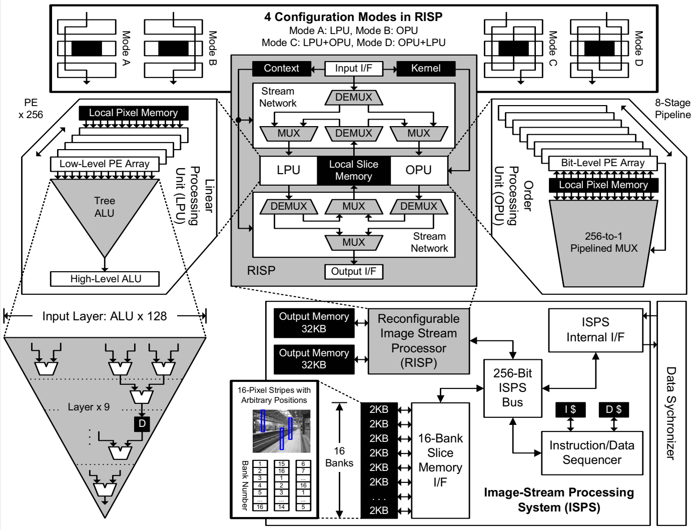


Figure 18.7.2: ISPS Architecture.

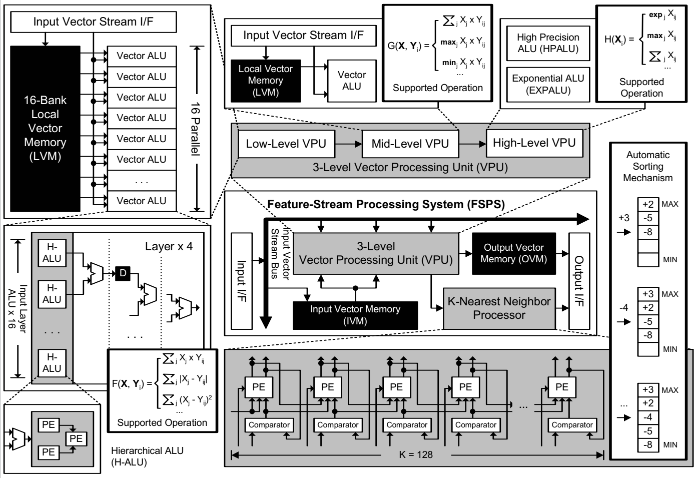

Input Vector Stream I/F
16-Bank Local Vector Memory (LVM)
Vector ALU
16 Parallel
Local Vector Memory (LVM)
Supported Operation
High Precision ALU (HPALU)
Exponential ALU (EXPALU)
Low-Level VPU
Mid-Level VPU
High-Level VPU
3-Level Vector Processing Unit (VPU)
Automatic Sorting Mechanism
Feature-Stream Processing System (FSPS)
Input I/F
Input Vector Stream Bus
3-Level Vector Processing Unit (VPU)
Output Vector Memory (OVM)
Input Vector Memory (IVM)
K-Nearest Neighbor Processor
Output I/F
H-ALU
Layer x 4
D
Input Layer ALU x 16
PE
Hierarchical ALU (H-ALU)
Supported Operation
Comparator
K = 128
MAX
MIN

Figure 18.7.3: FSPS Architecture (Machine-Learning Engine).

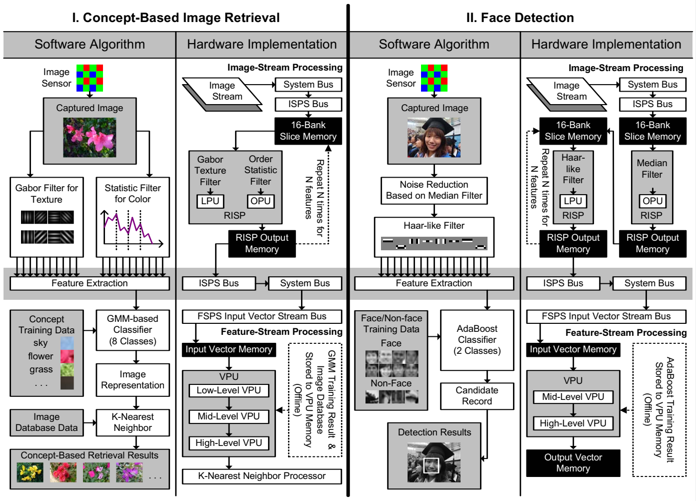
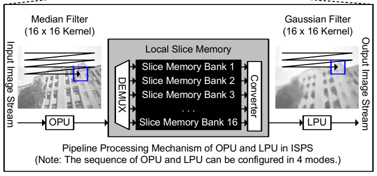


Figure 18.7.4: Example applications of the SASoC.

Performance analysis of some single operations (ISPS):
(The maximum frequency of ISPS is 200MHz.)

Operation Description	Processing Unit	Input Data Rate	Frame Rate (160x120)	
			ISPS	*Pentium C2D**
3x3 Edge Detector	LPU	1.7Gpixel/s	10106fps	*64fps*
5x5 Morphological Filter	OPU	4.7Gpixel/s	9823fps	*9fps*
7x7 Haar-like Filter	LPU	9.0Gpixel/s	9551fps	*21fps*
9x9 Median Filter	OPU	14.4Gpixel/s	9290fps	*0.7fps*
11x11 Sharpening Filter	LPU	21.0Gpixel/s	9040fps	*9fps*
13x13 Order Statistic Filter	OPU	28.6Gpixel/s	8799fps	*0.7fps*
15x15 Laplacian Filter	LPU	37.0Gpixel/s	8569fps	*5fps*
16x16 Gabor Filter	LPU	41.6Gpixel/s	8457fps	*5fps*
16x16 Gaussian Filter + 16x16 Median Filter	LPU + OPU	76.8Gpixel/s	7004fps	*0.4fps*
16x16 Median Filter + 16x16 Gaussian Filter	OPU + LPU			



Pipeline Processing Mechanism of OPU and LPU in ISPS
(Note: The sequence of OPU and LPU can be configured in 4 modes.)

Performance of some single operations (FSPS):
(The maximum frequency of FSPS is 200MHz.)

Operation Description	Class Number	Input Data Rate	
		FSPS	*Pentium C2D**
AdaBoost Classifier	2	3.2G dimension/s	*8.6M dimension/s*
Support Vector Machine (SVM): RBF Kernel*	2	3.2G dimension/s	*171M dimension/s*
Support Vector Machine (SVM): Linear Kernel*	2	51.2G dimension/s	*312M dimension/s*
Artificial Neural Network (ANN)	2	51.2G dimension/s	*384k dimension/s*
Gaussian Mixture Model-based Classifier (GMM)*	2 ~ 8	3.2G dimension/s	*69.2k dimension/s*

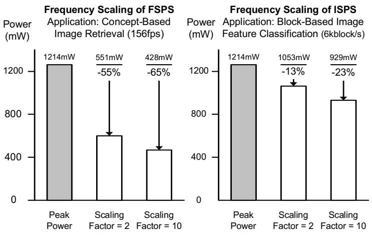
Operation Description	Dimension	Input Data Rate	
		FSPS	*Pentium C2D**
Mahalanobis Distance	1 ~ 16	0.4Gvector/s	*1.4kvector/s*
Euclidean Distance	1 ~ 16	0.4Gvector/s	*2.3Mvector/s*

Operation Description	Dimension	Input Data Rate (Database)	
		FSPS	*Pentium C2D**
K-Nearest Neighbor	1 ~ 16	0.2Gvector/s	*17kvector/s*
	17 ~ 32	0.1Gvector/s	*10kvector/s*
	33 ~ 48	67Mvector/s	*7kvector/s*
	49 ~ 64	50Mvector/s	*5kvector/s*
	65 ~ 80	40Mvector/s	*4kvector/s*
	81 ~ 96	33Mvector/s	*3kvector/s*
	97 ~ 112	29Mvector/s	*3kvector/s*
	113 ~ 128	25Mvector/s	*3kvector/s*

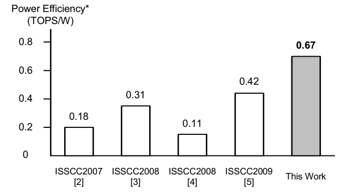
**Note 1: PC with Pentium Core 2 Duo CPU (2.83GHz) and 4GB RAM is used for performance comparison.*
**Note 2: The performance of SVM is measured based on LIBSVM (Chang and Lin, 2001).*
**Note 3: The computation of Input Data Rate for GMM: (Dimension Number per Vector) x (Class Number) x (Gaussian Number per Class) / (Total Time).*

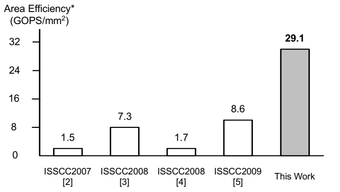
Figure 18.7.5: Performance analysis of the ISPS and FSPS.

Process	TSMC 90nm IP9M CMOS	
Die Size	$28mm^2$ (7mm x 4mm)	
Power Supply	Core 1.2V, I/O 2.5V	
Operating Frequency	3 Clock Domains Maximum 200MHz for Each	
Total Gate Count (2-Input NAND Gate)	3.00M gates (ISPS: 1.23M gates, FSPS: 1.76M gates)	
On-Chip SRAM	149KB (ISPS: 100KB, FSPS: 49KB)	
Peak Performance	814GOPS (ISPS: 512GOPS, FSPS: 302GOPS)	
Maximum Input Data Rate	ISPS	76.8Gpixel/s
	FSPS	51.2Gdimension/s
Power Consumption	1214mW (Peak)	
Power Efficiency	671GOPS/W	
Application Example (160 x 120 frame)	Concept-Based Image Retrieval: 156fps	
	Face Detection: 294fps	



Technology Scaling of Power:
(130nm to 90nm) $P_{90}=P_{130}x(C_{90}/C_{130})x(V_{90}/V_{130})^2=P_{130}x0.69x(1.2/1.2)^2=0.69$
(65nm to 90nm) $P_{90}=P_{65}x(C_{90}/C_{65})x(V_{90}/V_{65})^2=P_{65}x1.38x(1.2/1.0)^2=1.99$





**Note: The comparison is performed without on-chip memory power/area considerations.*

Figure 18.7.6: Chip features and comparisons.

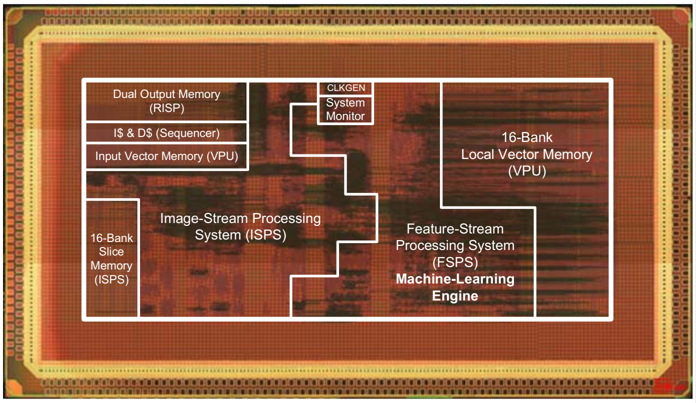

Dual Output Memory (RISP)
I$ & D$ (Sequencer)
Input Vector Memory (VPU)
CLKGEN
System Monitor
16-Bank Local Vector Memory (VPU)
16-Bank Slice Memory (ISPS)
Image-Stream Processing System (ISPS)
Feature-Stream Processing System (FSPS)
Machine-Learning Engine

Figure 18.7.7: Chip micrograph.