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 1280 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 76.8GB/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 28mm² 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 support and chip testing. This work is funded by National Science Council and TSMC.

References:
I. Concept-Based Image Retrieval

II. Face Detection

Software Algorithm

Hardware Implementation

Software Algorithm

Hardware Implementation

Figure 18.7.1: SASoC Architecture.

Figure 18.7.2: ISPS Architecture.

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

Figure 18.7.4: Example applications of the SASoC.

Figure 18.7.5: Performance analysis of the ISPS and FSPS.

Figure 18.7.6: Chip features and comparisons.
Figure 18.7.7: Chip micrograph.