B-Face: 0.2 mW CNN-Based Face Recognition Processor with Face Alignment for Mobile User Identification

Sanghoon Kang, Jinmook Lee, Changhyeon Kim, and Hoi-Jun Yoo
School of EE, KAIST, Yuseong-gu, Daejeon, Republic of Korea
E-mail: sanghoon_kang@kaist.ac.kr

Abstract

An ultra-low-power face recognition processor with binary weight convolutional neural network core and face alignment accelerator is proposed for high accuracy even with head pose variations. Binary convolution core with exhaustive input reuse and interleaved output memory access is proposed to minimize power consumption, resulting in 13.3 TOPS/W power efficiency. In addition, face alignment core with zero-aware pipelining is proposed to minimize external memory access. As a result, the face recognition system with maximum 48 frames-per-second throughput and 0.2 mW minimum power consumption is realized.

Introduction

Recently, face recognition (FR) is widely adopted to mobile devices as a measure for user identification, enabling user-specific services. The convolutional neural network (CNN) shows high accuracy in FR, so that a dedicated CNN FR processor was developed to reduce power consumption in battery-limited environment [1]. However, variance in head poses causes accuracy degradation for stand-alone CNN up to 10% [2], as shown in Fig. 1. Face alignment which converts the input face into its frontal face can enhance the recognition accuracy, but its software realization on mobile application processors consumes >1W power, greatly diminishing battery run-time. The necessity of the face alignment hardware was first raised in [3], but only simulated results were given. Therefore, a low-power FR processor with face aligning capability is necessary for high accuracy mobile FR systems.

The previous CNN processor [1] showed low-power consumption with 8-bit precision weights, but its power efficiency is low due to inefficient data reuse and transaction pattern among processing elements (PE) and its distributed memories. Recently, [4] showed the weight precision of CNN to be reduced down to 1-bit, without considerable sacrifice in classification accuracy. The 1-bit weight precision helps the chip implementation by reducing the hardware complexity drastically, but its accuracy is susceptible to degradation under face pose variations. Therefore, the face alignment plays a more important role in binary weight CNN FR. A binary-weight CNN FR with face alignment shows only ~1% accuracy degradation compared to [1], with just 326 KB of total weight size.

This work proposes an FR processor, with binary convolution core (BCC) and face alignment core (FAC) for ensuring high recognition accuracy under pose variations. For low-power consumption, 3 features are proposed: 1) binary convolution engine (BCE) with exhaustive input reuse, 2) pipelined partial sum accumulation engine (PAE) with interleaved output memory (OMEM) access, and 3) face alignment engine (FAE) with zero-aware pipelining.

Proposed FR Processor

Fig. 2 shows the overall architecture of the proposed FR processor. It consists of 2 separate cores: BCC and FAC. The 12-way parallel BCEs and PAE are integrated into BCC, alongside convolutional loading unit (CLU), 36 KB weight memory (WMEM) and 12 KB OMEM. The FAC is composed of 4 pairs of face memory (FMEM) and alignment engine (AE). The FAC generates the aligned face image through iterations of regression to locate 32 facial feature points [5]. The 128×128 aligned face image is recognized through CNN on the proposed BCC.

1. The BCEs with exhaustive input reuse (Fig. 3) generates 12 partial sums per clock cycle, each of which is a result of 64 multiply-and-mac (MAC) operations. The binarization of CNN weights changes the data pattern, from weight-oriented to feature-oriented. Since the entire binary weights of each convolution layer of the target FR CNN can be stored in 36 KB WMEM, input feature reuse becomes the first priority to minimize external memory access (EMA). In this paper, a dictionary based BCE is proposed to reuse input feature data exhaustively. The dictionary contains $2^{16}$ entries of binary weight convolution results, which are the all-possible outcomes of 4-way binary weight MAC operation. After generation of dictionary based on the 4 loaded input features (a, b, c, d), MAC results are indexed from the 4 binary weights. Compared to standard adders, proposed dictionary based MAC with 16 entries shows 54% energy reduction, with only 16% area overhead. Once input features are loaded, weights for the entire output channels are fed in consecutive clocks from 12×64-bit WMEM. Moreover, convolutional reuse of nearby pixels is covered by CLU [6]. As a result, the BCC maintains high utilization, achieving 14% higher FR MAC efficiency (OPS/# MAC units) compared to [1].

2. Pipelined PAE with interleaved OMEM access (Fig. 4) increases the convolution pipeline utilization. PAE is composed of 12 adders to accumulate 12×16-bit partial sums from BCEs. Due to the exhaustive input reuse of the BCEs, the partial sums from the subsequent clock cycle correspond to different output channels from the current ones. This causes an OMEM access overlap as shown in Fig. 4(b), reducing the pipeline utilization of BCEs down to 50%. To eliminate the bubble, OMEM is divided into 2, yet maintaining the total storage amount. PAE accesses the OMEMs in an interleaved order, increasing the pipeline utilization to 100%.

3. FAE with zero-aware pipelining (Fig. 5) generates the aligned face image, with low-power consumption and reduced EMA. Each FAE in FAC is paired with 2 KB FMEM. Only 2 FAE-FMEm pairs are activated for the facial feature detection; one for the tree search, and the other for the regression. On the other hand, all 4 pairs are activated for the image warping, with each pair allocated to different rows. The regression weight matrix requires 2.93 MB of EMA per frame, which is 9-times larger than that of binary weight CNN. Utilizing the sparsity of the binary vectors from the tree search, only the indices of 1 are flagged, and only the selected rows are fetched from the external memory, achieving 93.8% EMA reduction. Moreover, tree search, weight fetch, and accumulation form a pipeline in the unit of trees, hiding memory access latency, and increasing the throughput by 56%.

Implementation Results and Conclusions

The proposed chip is fabricated in 65 nm CMOS process, as shown in Fig. 7. It occupies a small area of 3.2mm$^2$ with a total of 56 KB on-chip SRAM. The processor can operate from 5 MHz to 100 MHz clock frequency, with 0.66 V to 1.2 V supply voltage. The recognition of the binary weight CNN with face alignment shows 96% accuracy of LFW [7] at equal-error-rate. Face recognition with alignment runs in 1 frame-per-second (fps) throughput at 5 MHz and 0.66 V, with average power 0.2 mW. The maximum throughput is 47.6 fps, and the peak performance of the BCC is 154 GOPS. Table 1 shows the comparison with face recognition processors [1, 8], and deep learning processors [1, 9-10]. The proposed chip shows the state-of-the-art peak power efficiency of 13.3 TOPS/W. In conclusion, binary CNN FR processor is realized with face-alignment showing robustness to head pose variance, with maximum 48 fps throughput and minimum 0.2 mW power consumption.
Reference


Fig.1 Proposed face recognition system compared with previous FR system.

Fig.2 Overall architecture of the proposed FR processor.

Fig.3 Binary convolution core: (a) BCE architecture, (b) data pattern change with binarization, (c) data feeding & reuse pattern.

Fig.4 Partial sum management: (a) OMEM & PAE architecture, (b) convolution pipeline utilization.

Fig.5 Face alignment core: (a) FAE architecture with workload mapping, (b) zero-aware pipeline.

Fig.6 Chip photograph and performance summary.

Fig.7 Measurement results.

Table 1. Comparison table.