A 0.76mm² 0.22nJ/Pixel DL-assisted 4K Video Encoder LSI for Quality-of-Experience over Smart-Phones

Tsu-Ming Liu, Chang-Hung Tsai, Tung-Hsing Wu, Jia-Ying Lin, Li-Heng Chen, Han-Liang Chou and Chi-Cheng Ju

Mediatek Inc., Hsinchu, Taiwan
E-mail: mingle.liu@mediatek.com; cc.ju@mediatek.com

Abstract
This paper proposes the world’s first deep learning (DL)-assisted video encoder LSI fabricated in a 10nm process with a core area of 0.76mm² to integrate quad-core DL accelerators and 4K×2K H.264/H.265 video standards. A visual-contact-field network (VCFNet) DL model is newly designed to predict human focus information for extraordinary reducing the encoding complexity, leading to 82.3% of power reduction. Moreover, input channel reduction and layer merging approaches reduce VCFNet complexity by 69%. Operated at 0.9V and 504MHz, the proposed DL-assisted 4K video encoder LSI consumes 56.54mW to achieve 0.22nJ/pixel of energy efficiency, cutting 0.1-14nJ/pixel compared to conventional designs [1]-[3].

Introduction
Smart phones are gaining popularity these days while a large amount of videos are generated everyday and transmitted over the network. Current voice/video applications would be able to retain acceptable quality of experience (QoE) but the power consumption is one of the most important key influential factors on the overall perceived quality of smart-phones.

Several power-efficient video encoder LSIs [1]-[3] are reported to date, but they cannot adjust quality by QoE. In this work, QoE quality is achieved by concentrating on human’s visual contact for specific field of view (VCF): if more bits (or high-quality coding-tool) are assigned to the visual contact region, we can improve the subjective qualities with the same transmission bandwidth under the low bit-rate mode. For low-power mode, we can transmit a video of the same subjective visual quality by adopting low-complex coding tool in non-visual contact region so as to reduce the encoder power.

Fig 1. Proposed DL-assisted video recording for streaming system.
Fig. 1 depicts the DL-assisted video recording for video streaming system. In particular, the VCFNet is implemented by a fully convolutional neural network with 18-layer which consists of 2 basic feature layers, 5 VCF feature blocks (VFBs), and 2 VCF detection layers. And input channel reduction and layer merging approaches are applied to reduce the network complexity by 69%. In low-power mode, video encoder adjusts the coding tool and quality controller so as to reduce the complexity on non-visual contact region, leading to 82.3% of power reduction in 4K×2K resolution and 30fps while maintaining the same mean opinion score (MOS) in terms of subjective visual quality [5].

DL-assisted Video Encoder System Architecture
A. Block Diagram
Fig. 2 shows the system block diagram of the proposed video encoder processor with VCFNet DL accelerator, VCFMap engine, and video encoder. Encoding modes and parameters are adaptively adjusted via quality and coding tool tables (Q-Table and T-Table). The intermediate data in VCFNet and video encoder are shared in MMU in a time interleaving way. The main CPU is tri-cluster deca-core architecture [4] and ISP is dual 14-bit architecture and supports 28MPixel at maximum.

B. VCFNet DL Accelerator
It is evident that deep learning approaches provide superior results for recognition problems [6]. In this work, the proposed VCFNet DL model is trained to predict human’s visual contact by a large scale of training dataset, which takes the ISP processed image as input and outputs the visual contact field map of input image (VCFMap, 0-255 gray level, higher level means higher frequency of visual contact). To achieve 4K@30fps real-time requirement with a deeper (18-layer) network, a VCFNet DL accelerator is designed in Fig. 3. It comprises of four deep-learning processors and is connected to
model and feature map buffers. By adopting input channel reduction and mean/bias layer merging scheme, both memory bandwidth and computation are saved, leading to 69% of complexity reduction. Moreover, a dynamic VCF feature selection method is proposed to adaptively adjust the depth of VCFNet in each macro-block during video coding. From the experiments, 55% power consumption is reduced compared to computation in fine-layer under the same MOS quality.

C. VCFMap Engine

To further improve the power efficiency of VCFNet DL accelerator, a VCFMap engine is designed in Fig. 4(a). A 2D processing element (PE) array calculates the statistics of VCF map including active area and map difference between current and previous VCF map. In a live video call scenario, this map post-processing can further reduce the system power by 56% since the map difference is small in a slow motion sequence. The VCFMap engine not only reduces the VCFNet power but adjusts video encoder quality and coding tools through Q-Table and T-Table during video encoding in Fig. 4(b). Higher level in VCF map indicates the visual contact and therefore allocates higher complexity and more bits than lower level in VCF map.

Implementation Results

This chip is fabricated at a 10nm FinFET process with 0.76mm² area, and the chip micrograph is shown in Fig. 5. To clarify the system performance and visual quality, a system evaluation board is designed with a 4.7" WQXGA (2560×1600) display panel. In low bit-rate mode, a VGA 250kbps sequence is tested. VCFNet indicates the field of view for visual contact and encoder applies a higher bits in visual contact region and lower bits in non-visual contact region. In Fig. 6(a), a subjective visual quality can be greatly improved from 2 to 4 in MOS [5] measurement. In low power mode, 4K@30fps video, Basketball Drive for VCFNet and H.265 video encoder, is tested under 504MHz of working frequency. And 56.54mW of power dissipation, 0.22nJ/pixel of energy efficiency, is measured. Compared to state-of-the-art designs [1]-[3], this works cuts 0.1-14nJ/pixel of energy consumption in Fig. 6(b).

References