applied sciences

MDPI

Article

Could We Realize the Fully Flexible System by Real-Time Computing with Thin-Film Transistors? †

Qin Li ‡, Zheyu Liu ‡, Fei Qiao *, Qi Wei * and Huazhong Yang

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China;
li-q16@mails.tsinghua.edu.cn (Q.L.); zy-liu15@mails.tsinghua.edu.cn (Z.L.); yanghz@tsinghua.edu.cn (H.Y.)

* Correspondence: qiaofei@tsinghua.edu.cn (F.Q.); weiqi@tsinghua.edu.cn (Q.W.); Tel.: +86-10-6279-6637 (F.Q.)

† This paper is an extended version of the papers published at the International Symposium of Circuit and System (ISCAS), Baltimore, MD, USA, 28–31 May 2017; IEEE Computer Society Annual Symposium on VLSI, Pittsburgh, PA, USA, 11–13 July 2016.

‡ These authors contributed equally to this work.

Received: 31 October 2017; Accepted: 24 November 2017; Published: 27 November 2017

Abstract: Flexible electronic devices, such as the typical thin-film transistors, are widely adopted in the area of sensors, displayers, wearable equipment, and such large-area applications, for their features of bending and stretching; additionally, in some applications of lower-resolution data converters recently, where a trend appears that implementing more parts of system with flexible devices to realize the fully flexible system. Nevertheless, relatively fewer works on the computation parts with flexible electronic devices are reported, due to their poor carrier mobility, which blocks the way to realize the fully flexible systems with uniform manufacturing process. In this paper, a novel circuit architecture for image processing accelerator using Oxide Thin-film transistor (TFT), which could realize real-time image pre-processing and classification in the analog domain, is proposed, where the performance and fault-tolerance of image signal processing is exploited. All of the computation is done in the analog signal domain and no clock signal is needed. Therefore, certain weaknesses of flexible electronic devices, such as low carrier mobility, could be remedied dramatically. In this paper, Simulations based on Oxide TFT device model have demonstrated that the flexible computing parts could perform 5 × 5 Gaussian convolution operation at a speed of 3.3 MOPS/s with the energy efficiency of 1.83 TOPS/J, and realize image classification at a speed of 10 k fps, with the energy efficiency of 5.25 GOPS/J, which means that the potential applications to realize real-time computing parts of complex algorithms with flexible electronic devices, as well as the future fully flexible systems containing sensors, data converters, energy suppliers, and real-time signal processing modules, all with flexible devices.

Keywords: flexible electronics; thin-film transistors; image signal processing; machine learning; analog-to-information processing; physical computing

1. Introduction

With the advantages of transparency, softness, biocompatibility, etc., flexible electronic devices have been widely used in displaying and sensing fields [1]. For example, the flexible electronic devices have put the curved screen [2] and the electronic skin [3] into practice. Meanwhile, state of the art has done a lot of proven researches in the flexible self-energizing (solar, etc.) [4] and flexible antennas [5]. In some applications, like heart care, where biocompatibility is extremely vital, realizing a fully flexible system that including flexible sensing, computation, and display are necessary for its biocompatible physical properties.

As shown in Figure 1, the innovative fully flexible system architecture, where the works on various types of flexible interfaces toward bio-tissue [3], flexible analog-to-digital converter (ADC) [6],

memory [7], sensors, and display [1,2] have been reported, however, the computation based on flexible devices is still an unsolved problem for the fully flexible system, especially for the complex algorithms with real-time processing requirements. The lower carrier mobility of the devices themselves and device deviation that is caused by immature manufactory make it impossible to implement the high-frequency and high-resolution digital processing system on flexible technology. Besides, high-speed, high-resolution flexible ADC and Digital Signal Processing (DSP) are needed if the traditional digital processing flow would be adopted, which is also energy hungry. In order to make up for the shortcomings of flexible computing, Yoon [8] has tried to combine the flexible devices with the silicon-based devices, where the complex computation part is done by silicon-based integrated circuits. Nevertheless, this kind of pseudo-fully-flexible system brings a lot of problems inevitably. Silicon is too hard to be applied in the biocompatible application and the interfaces between two devices are costly due to the totally different manufacturing process. Kris [9] proposes an 8-bit general-purpose microprocessor that is based on thin-film transistor (TFT) devices, which could only perform simple algorithm, such as logic, arithmetic, and bit shift functions at a maximum frequency of 2.1 kHz. To sum up, the performance of TFT based digital circuit is not comparable with conventional silicon-based technology and is difficult to support the computation intensive real-time applications, such as the prevailing image signal processing and machine learning. So could we realize the fully flexible system by real-time computing with thin-film transistors?

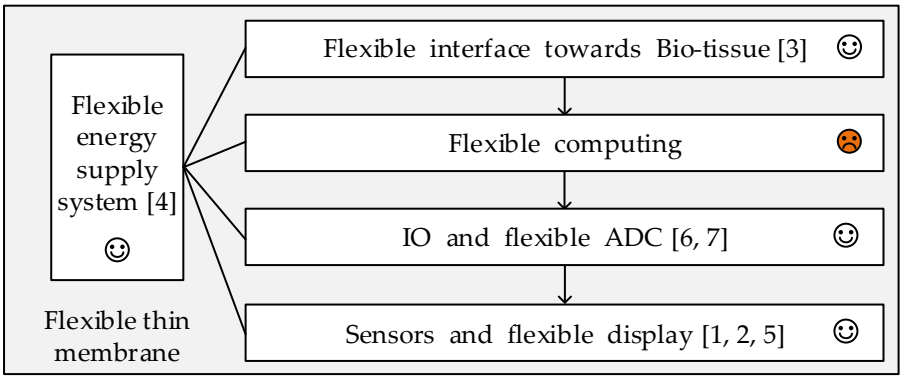

Figure 1. Innovative fully flexible architecture.

Analog circuits are well known as more energy-efficient and having a higher computing speed than their digital counterparts, for the clock-free processing mode and lower resource consumption. Analog accelerators for signal processing also have recently achieved the breakthrough in various fields [10], but analog TFT circuit is rarely reported due to the device deviation and component mismatch, which are also the major disadvantage of the flexible device. However, in some fault-tolerant applications, the deviation may not be critical. For example, in feature extraction algorithms of image processing, such as Difference-of-Gaussian (DoG) [11] and Histogram of Oriented Gradient (HOG) [12], the feature descriptors are generated from local or global statistical information, in which the exact original data does not make much sense. Moreover, in machine learning algorithms, such as multi-layer perception (MLP) or convolutional neural network (CNN), the mismatch, and as such other distortions will be absorbed in the trained classifier [13]. We proposed the analog accelerator for image pre-processing that based on the fully flexible system [14,15]. Analog-to-Information processing (AIP) architecture was introduced to implement the typical neural network MLP [16]. We extended this paper to a fully flexible processing system that includes fault-tolerant pro-processing and computation-intensive processing for image classification. High-speed and energy-efficient image processing based on flexible devices is realized, which is the key to implement the fully flexible system. Besides, the error caused by process variation, temperature, etc. is analyzed, and fault-tolerance analysis of system is presented in this paper.

In the following parts of the paper, Section 2 will introduce the properties of flexible electronic device and the fault-tolerant image processing algorithms. The proposed analog accelerator for

image pre-processing and AIP architecture for image classification are presented in Section 3. The circuit design to solve the problems of flexible computing is shown in Section 4. The analysis of error and the ways to eliminate process variation are discussed in Section 5, and Section 6 concludes the paper.

2. Materials and Methods

2.1. Flexible Electronic Device

Table 1 shows the comparison of characteristics among CMOS (silicon-based MOSFET), A-Si:H-based thin-film transistor (TFT) and A-oxide-based TFT [18]. With the excellent physical properties including transparency and soft, flexible devices are irreplaceable to be applied in the fully flexible system. In addition, due to the low manufacturing temperature and low cost, the manufacturing cycle and price of flexible devices are much smaller than CMOS. However, the process variation of the flexible devices is large than CMOS due to the immature process. As the previous works report [17,18], the carrier mobility of A-oxide TFTs is 10–50 cm^2/Vs, which is an order of magnitude smaller than that of CMOS but still large than that of A-Si:H TFTs. That is, it is difficult to realize high-speed, high accuracy signal processing with such flexible devices.

Table 1. Comparisons among different devices [17,18,19].

Index	CMOS	A-Si:H TFTs	A-Oxide TFTs
Carrier Mobility (cm^2/Vs)	480–1350	<1	10–50
Transparent	×	√	√
Soft	×	√	√
Biocompatible	×	√	√
Process Variation	Small	Medium	Large
Cost	High	Medium	Low
Manufacture Temperature	High	≅110 °C	Room temperature

A-oxide TFTs technology provides 3-terminal thin-film transistors, with a typical channel length of 5 μm from the In-Ga-Zn-O system [18]. The substrate of this technology is plastic foil so that the circuits are flexible and even rollable. The electronics behavior of Oxide TFT transistors is quite similar to the crystalline silicon-based transistors. Thus the model adopted in our simulation is like MOSFET models but the model parameters are quite different. During the simulations, the mobility data adopted in the oxide TFT model (In-Ga-Zn-O) are around 10 cm^2/Vs, which could be referred to the experimental data from some TFT device research literature, such as [18], without loss of generality.

2.2. Fault-Tolerant Image Processing Algorithm

The realization of visual perception with flexible devices is promising and essential, which is also critical for greatly broadening applications of the flexible devices. Generally, image classification tasks are composed of pre-processing and feature processing phases, which are used to enhance the features extraction and classification, respectively. As for the pre-processing phase, the widely used Difference-of-Gaussian (DoG) algorithm [11] is adopted to extract and enhance edges of pictures. Additionally, as for the classification phase, multi-layer perception (MLP) [20] is selected because it contains the common multiply-accumulate units and non-linear activation units of neural network, without loss of generality.

Besides, the computation burden of the Gaussian convolution is higher and the mask size is larger with the increasing resolution of image, and it becomes a tough mission for digital processors, because the same multiplication and addition operations are performed on every single pixel and with neighbor pixels. Moreover, neural network is also a computation-intensive task, for its large number of computations between every two layers, which could not be performed on the general-purpose processors with real-time processing. Therefore, the DoG and MLP algorithm are

sufficient to verify our architectures, which solves the contradiction between low mobility, low stability device versus complex computation task.

Moreover, the fault-tolerant capability of the selected algorithms could tolerate the error that is introduced by unstable devices and analog computations. The feature descriptors of DoG are generated from local statistical information, in which the exact original data does not make much sense; and, the training process of neural network can gradually eliminate the error that is caused by imperfect computation. Details of the DoG and MLP are introduced as follows.

2.2.1. Pre-Processing Algorithms

Pre-processing algorithms are very important step in image processing including edge enhancing, features extracting, etc. Among them, the Difference-of-Gaussian (DoG) [11] is a widely used algorithm to extract and enhance edges. Actually, the human retina can be considered as a retina kernel filter whose model based on the DoG, which, thus, is an appropriate method that is applied to novel flexible computing.

The DoG is calculated by decreasing the two Gaussian functions with different scales. Assume that the input image is $I(x,y)$, the Gaussian kernel function is $G(x,y,\sigma)$ and output image processed by Gaussian convolution noted as $\mathrm{M}(x,y,\sigma)$, where x and y are horizontal and vertical coordinates in pixels space, and σ is the coordinate in scale space. Then, we have:

$$M(x,y,\sigma) = I(x,y) * G(x,y,\sigma) \quad (1)$$

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

The difference of one image filtered by $\sigma 1$ and $\sigma 2$ is:

$$D(x,y,\sigma 1,\sigma 2) = M(x,y,\sigma 1) - M(x,y,\sigma 2) \quad (3)$$

The feature descriptors here are generated from local statistical information, in which the exact original data does not make much sense. Thus, noise or fault in the appropriate range can be ignored. Above all, Gaussian convolution is a most common and critical operation in the DoG. It has been demonstrated by Lindeberg that the only possible scale-space kernel is the Gaussian function among a variety of reasonable assumptions [11,21]. Kang et al. [22] proposed a Gilbert Gaussian circuit, whose I–V characteristic is very close to Gaussian function. Based on this circuit, we construct an architecture to implement Gaussian convolution in the analog domain, and it is chosen to demonstrate the computation potential of Oxide TFT technology.

2.2.2. Classification Algorithms: Feature Processing

Multiple Layer Perceptron (MLP) [20] is the typical and classical algorithm of artificial neural networks (ANN) due to its computation units is common in the ANN. It is also widely used in the image classification and image fitting fields. In general, MLP mainly contains multiply-accumulate units and non-linear (NL) units.

As shown in Figure 2a, is the typical structure of three-layer perceptron. The input layer contains the data from the image sensors that have been pre-processed. Once the input data are loaded, then the weights of the first layer will be multiplied with the input data and the results are accumulated to the data of hidden layer. Next, the data of hidden layer will be processed by non-linear algorithm and then working out the results of the second layer. The non-linear used here is the sigmoid function, whose curve is shown in Figure 2b, and the function expression is:

$$\mathrm{Sigmoid}(x) = \frac{1}{1 + \mathrm{e}^{-x}} \quad (4)$$

After the multiply-accumulate operation of the output layer, the classification results are sent to the back propagation (BP) units to update weights. Larger size of hidden layer means higher classification accuracy; therefore, it is an excellent solution to balance accuracy and efficiency in

complex tasks. Moreover, the training process of neural network can gradually eliminate the error that is caused by imperfect computation.

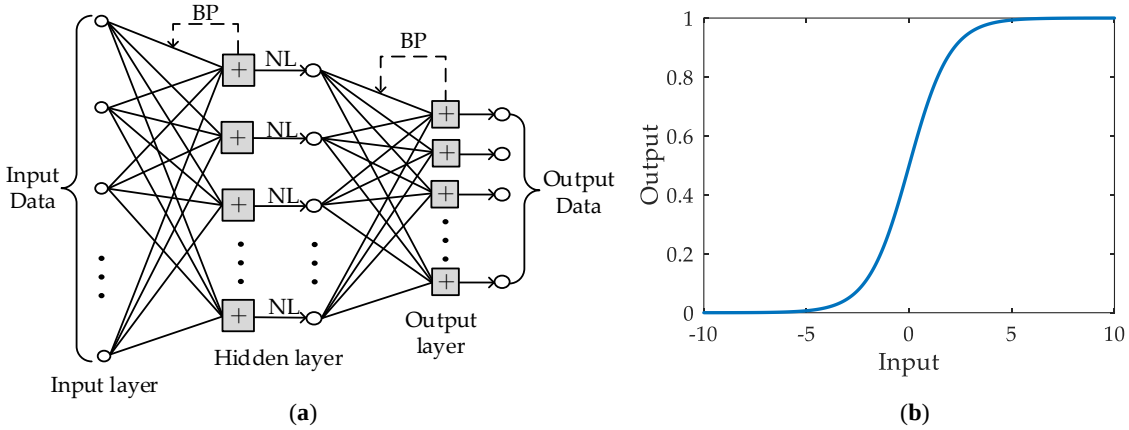

Figure 2. (**a**) Three-layer perceptron; (**b**) Curve of non-linear (NL) unit.

2.3. Analog-to-Information Processing Method

Neural network is a compute-intensive algorithm with amounts of input data and parameters. It is necessary to reduce the price of computation and make it more suitable for flexible circuits. "Let physics do compute" [23] is a new approach to deal with massive computational problems in the post-Moore era. It reveals the intrinsic physical characteristic of the electronic devices to enhance their functionality. That means that some calculations could be done directly on the original waveform but not on the bit streams. Based on this theory, a new computational paradigm named "analog-to-information processing" is proposed in this work. In this paradigm, some calculations and transform could be done right after the original analog data is obtained by sensors (shown in Figure 3). For example, in [24], active resistor network or MOSFET network is directly attached to the COMS image sensor's output, performing ultra-fast and energy-efficient Gaussian convolution on sensed images. However, resistors are hard to fabricate via the adopted Oxide TFT technology, and it is so rigid that we cannot modify the network parameters once the design is finalized. Therefore, we compose an oxide transistor network to replace the resistor network. The transistors are much more flexible than the resistors because of the programmability. Taking the advantage of cost-efficient and easy fabrication via Oxide TFT technology, a profusion of functional transistors that can be integrated on the large area of plastic foils. This fact allows for highly parallel computation in the analog domain, reaching higher performance than conventional digital architectures.

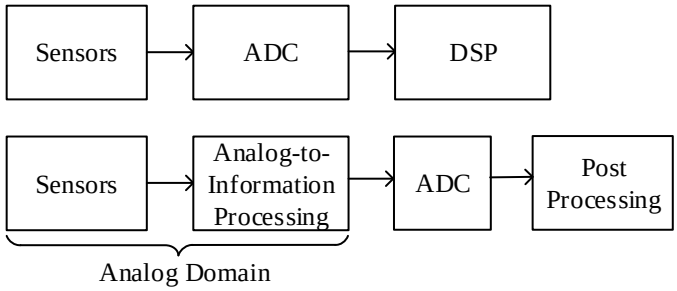

Figure 3. Tradition perception signal processing versus analog-to-information processing method.

3. Architecture and System Overview

As shown in Figure 4, we apply the AIP architecture on flexible electronic to realize energy-efficient and high-speed image processing. It is widely accepted that analog circuits are faster and more energy-efficient compared with digital circuits. Therefore, we make full use of the advantages of the analog domain processing to solve the low carrier mobility problem of flexible

devices. Although the unstable flexible devices and analog circuit will introduce noise, the algorithms we use have the fault-tolerant characteristics.

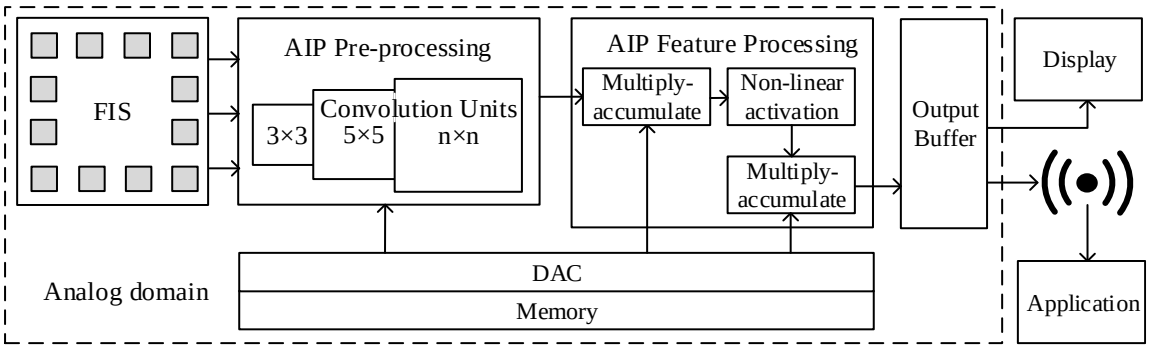

Figure 4. The model of analog-to-Information processing (AIP) architecture in fully flexible system.

The overall system architecture is illustrated in Figure 4. This architecture is aimed at realizing a fully flexible system, including sensor, memory, data converter, and the most important signal processing. The image sensor, as we mentioned before, is supposed to be the active pix sensor (APS) architecture to acquire current-mode pixel output. The outputs of the pixel are directly attached to the inputs of the Gaussian convolution unit. The input voltages are generated from on-chip memory and digital-to-analog converter (DAC), thus we can change the σ value according to the application requirement. The Gaussian convolution unit is organized as convolution mask, whose size is determined by the maximum possible σ value. The number of Gaussian convolution unit integrated into the image sensor decides the calculation parallelism, which will be analyzed later. But it also depends on the constraints of chip area, wire layout, and available power, so it requires careful planning.

The results of Gaussian convolution are sent to the AIP feature processing unit to do features extraction and classification. As presented in Section 2.2.2, input image that has been pre-processed is directly attached to the inputs of the multiply-accumulate units. After the operations of MLP, output features could be buffered, and read out for display or other application, under the control of address and bus controller. In this architecture, the sensing and computing are both done in the analog domain. There is no clock participating in the convolution process so that the calculation time is only related to the settling time of the circuit. In this way, the image processing could implement in real time using the relatively much slower Oxide TFT devices.

4. Circuit Design

4.1. The Basic Circuit Unit Design of Gaussian Convolution

According to Equations (1) and (2), Gaussian convolution basically consists of Gaussian multiplication and addition operations. In this presented circuit, the multiplication is implemented by Gilbert Gaussian multipliers. Due to the current mode output, the addition operation is simply realized by connecting the multipliers' outputs together, according to Kirchhoff's current law. The Gaussian convolution unit is illustrated in Figure 5 [15], where the numbers (from 1 to 3) means the different spread factors. According to [21], it is easy to prove that the Gilbert Gaussian circuit works as a multiplier between the tail current and the Gaussian function, using the exponential transfer characteristic of the subthreshold region of transistors. Thus, the tail current could be replaced by the pixel output to perform the Gaussian multiplication on pix data. Because of the current-mode input, the active pix sensor (APS) architecture could be used as the front end. The differential voltage input of the multiplier is used to tune the scale or the spread factor σ of the Gaussian kernel function. Once the transistors' parameters are decided, a default σ is determined. In this work, the multiplier's input-output relation is simulated to be:

$$I_{\text{out}} = I_{\text{in}} \times \exp(-6.6 \times \Delta V^2) \quad (5)$$

If we need an arbitrary σ to perform multi-scale filtering, we only need to calculate the input ΔV by:

$$\Delta V = \frac{1}{\sqrt{13.2}\sigma} \cdot x \quad (6)$$

By substituting Equation (6) to (5), it is easy to see the relation transform to:

$$I_{\text{out}} = I_{\text{in}} \times \exp(-\frac{x}{2\sigma^2}) \quad (7)$$

The x is an integer in $[-n, n]$, representing the corresponding position in one-dimension convolution mask. Its value range is in proportion to the mask size. This transformation is easy to be generalized to the two-dimensional (2-D) situation. When the factor σ and the position of the multiplier are decided, the input voltage ΔV can be calculated via Equation (6). Therefore, it is easy to implement on circuit via on-chip memory and DAC, which have been demonstrated on flexible technology. The output is read out in voltage to match the input of MLP.

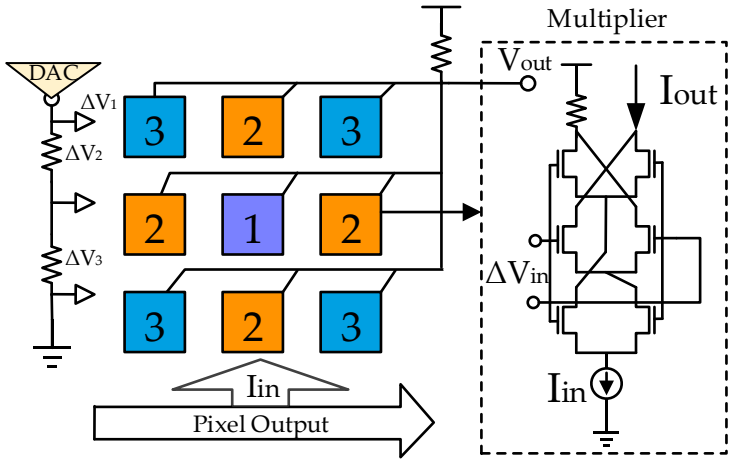

Figure 5. Gaussian convolution unit in analog signal domain.

4.2. The Basic Circuit Unit Design of MLP

As mentioned in Section 3, input signals that from the flexible pre-processing (FPP) circuit are connected with multiplying circuit of the first layer directly in the form of voltage, where the Gilbert multiplier [25] is applied because of its high linearity. Because the weights of MLP have been trained off-line, it is load beforehand on the differential ports of multipliers from the on-chip memory. That is, once the data is loaded, results of multiplier-accumulator (MAC) in the first layer, which is contained in the differential voltage of the output, can be calculated directly. Without digital clock, the computing circuit's performance has direct correlation with the setting time of the circuit path. As shown in Figure 7, addition in analog circuit is achieved by directly connecting the input nodes, nevertheless, amounts of input nodes would introduce a high voltage output node (*V*a or *V*b) that causes the input MOS working in the linear region. Therefore, the differential outputs of multiplier are subtracted to obtain single-end output first by current mirror. When considering the complexity of circuit design, ten multipliers with one current mirror to form a processing unit (from $\text{PU}(1,1)$ to $\text{PU}(1,m)$), which is shown in Figure 6. For the dataset MNIST, the dimension of input image is 28 × 28, thus the value of m in $\text{PU}(1,m)$ is $\left\lceil \frac{28 \times 28}{10} \right\rceil = 79$. Then, the single-end outputs of multipliers are accumulated to the input of hidden layer. When considering the tradeoff between power consumption and accuracy, the size of hidden layer is 50, which means that the value of n in PU_n is 50.

After the computation in the fully connection layer, intermediate results, or usually called hidden layer, are processed by nonlinear function Sigmoid. As shown in Figure 7, there are two transistors in sigmoid structure, whose grids connect input voltage and differential of drains represent output voltage, which can be written as [21]:

$$V_{\mathrm{od}} = I_{\mathrm{SS}} \times \frac{2\mathrm{R}}{1 - \exp(-\frac{2V_{\mathrm{id}}}{nV_{\mathrm{T}}})} \quad (8)$$

where V_{id} is the differential input voltage between two gates of transistors, R is the resistance, I_{ss} is the tail current, and n, V_T is the subthreshold-slope parameter and thermal voltage. Then, the signals come into the second MAC, where multiplying unit in the output layer is the same structure with the first one except the parameters of the circuit, for instance, source voltage, resistance value, and breadth length ratio of each transistor and tail current. At last, the computation results are sent to post-processing unit for further application.

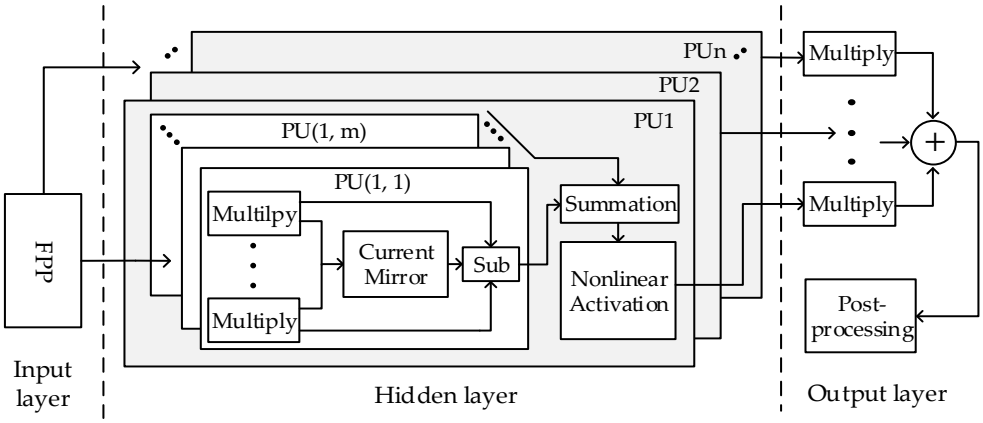

Figure 6. The computation process of multi-layer perception (MLP) based on flexible devices.

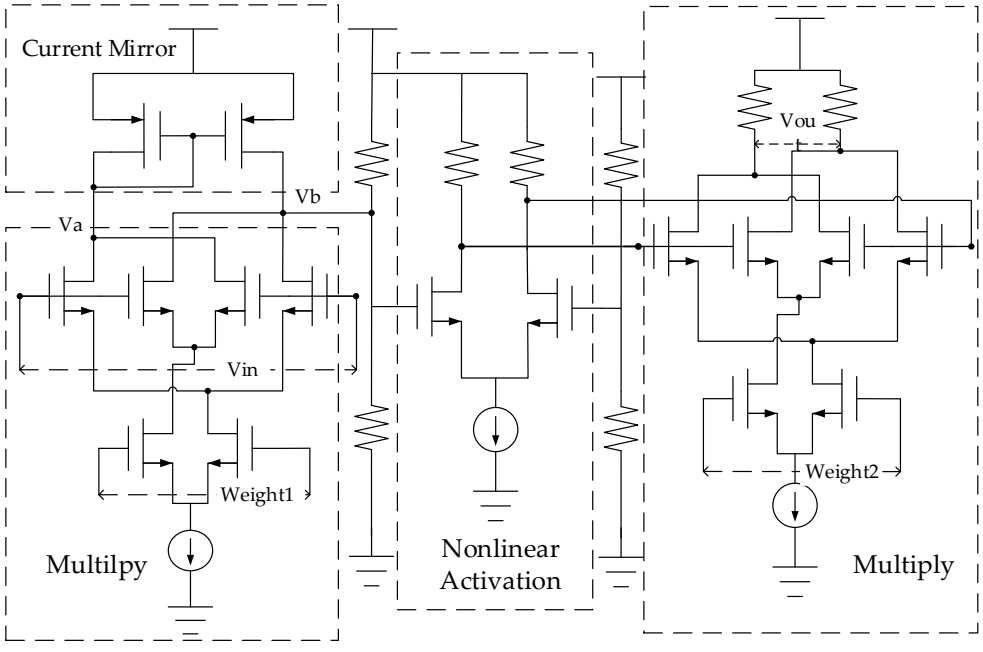

Figure 7. The circuit implementation of multiplier-accumulator (MAC) and Non-linear activation.

5. Simulations and Analysis

5.1. Verification of Functional Correctness

To determine whether our design could perform convolution in the correct way, we choose two continuous scale coordinates in scale space: $\sigma 1 = 1.5450$ and $\sigma 2 = 2 = 1.9466$. Then, the corresponding Gaussian convolution units are constructed based on the Oxide TFT model. A 128 × 128 test image is used as the simulated sensor output. The circuit simulation is done by Cadence; meanwhile, the

comparison digital result is calculated by MATLAB. The images after Gaussian filtering are shown in Figure 8a [15]. It is a little hard to tell the difference between the circuit output and the digital results, so we use PSNR to measure the accuracy of the circuit computed result, taking the software results as precise data. Calculating formulas for PSNR, are shown as following:

$$\text{PSNR} = 10\log_{10}\left(\frac{255}{\text{MSE}}\right) \tag{9}$$

$$\text{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}|I(i,j) - S(i,j)|^2 \tag{10}$$

where I represents the circuit-output image, and S is the precise data. The PSNR of two scales is 54.1 dB and 50.7 dB respectively. The reason why the larger σ has the lower accuracy is that the corresponding mask size is enlarged when the σ increases. The larger mask size causes more information lost on the edge of the image. It is easy to see that the edge of circuit output images appears to be black, while the software outputs have more smooth edges because it is easy to compensate the edge pixel in the software algorithm. However, it is a relatively acceptable result for image filtering. This simulation demonstrated the functionality of the proposed Gaussian convolution unit design.

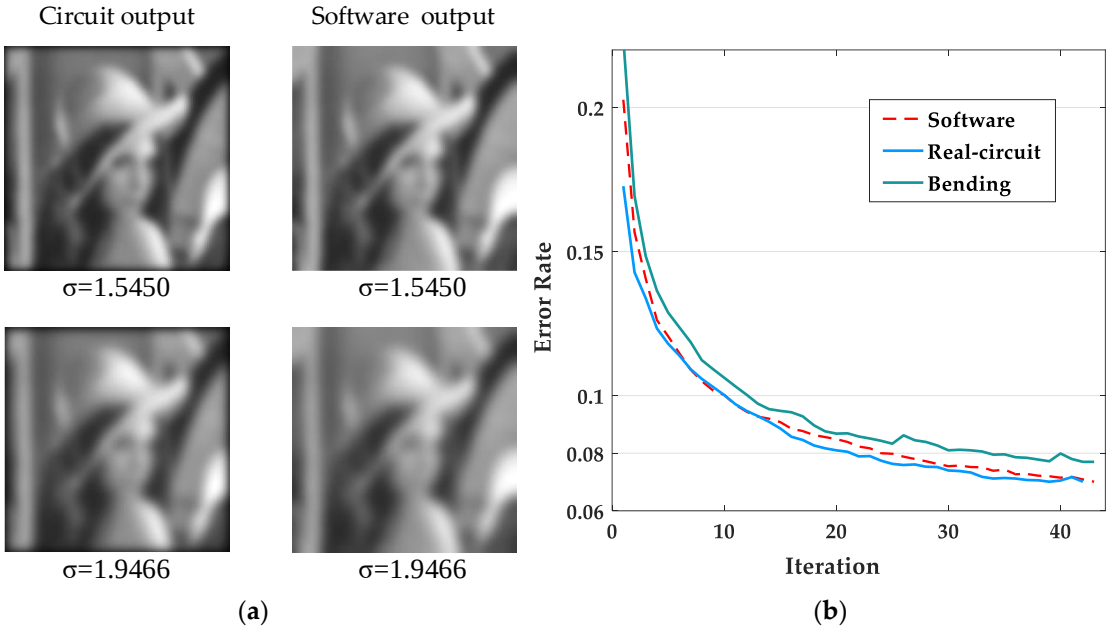

Figure 8. (**a**) Comparison of flexible computing circuit and software outputs of implementing convolution algorithm; (**b**) Classification error rate of MLP.

As shown in Figure 8b, software and flexible circuit simulation results are compared for the capability of classifying input image, where the typical datasheet MNIST with the size of 28 × 28 is used. The circuit simulation is done by HSPICE level 62 model for oxide TFT, and the software simulation is based on MATLAB. With the same network size, the error rates of software and circuit implementation both converge to the same value of 7.01%. Flexibility is an important property of Oxide TFT. According to [18], the mobility of oxide TFT decreases by 15.7%, with the device being bend at R = 30 mm. In order to evaluate the system performance after the bending of oxide TFT, the accuracy simulation with the mobility decreasing by 20% is given as Figure 8b. The error rate with the bending oxide TFT converges to the value of 7.7%, which is only 0.69% lower than the circuit implementation at the same network size, and still has a fast convergent trend. That is, although the analog circuit and the bending of oxide TFT introduce variations to computation, the MLP eliminates the error by training.

5.2. Performance Analysis

5.2.1. Settling Time

The following simulations are based on the HSPICE level 62 model for oxide TFT. As we have stated above, the speed of calculation basically depends on the settling time of the Gaussian convolution unit. We simulate the settling time for several units with different scales and different step-input current sources. The normalized settling time curves are shown in Figure 9a. The settling time of Gaussian convolution circuit is about 300 ns. It is equivalent to perform the convolution operation at a frequency of 3.3 MOPS/s, which is much faster than the conventional digital circuit driven by kHz-level clock on TFT [9]. As to the feature processing, the settling time of flexible MLP is about 100 μs, which means that the classification speed is up to 10 k frames/s, and is totally enough for real-time image processing with the speed of 30~100 fps. This result is really significant because it proves that, by exploiting the physical characteristic, the slow devices could exceed its limitation to reach a settling-time-equivalent calculation speed. These, in fact, demonstrate our thoughts of analog-to-information processing.

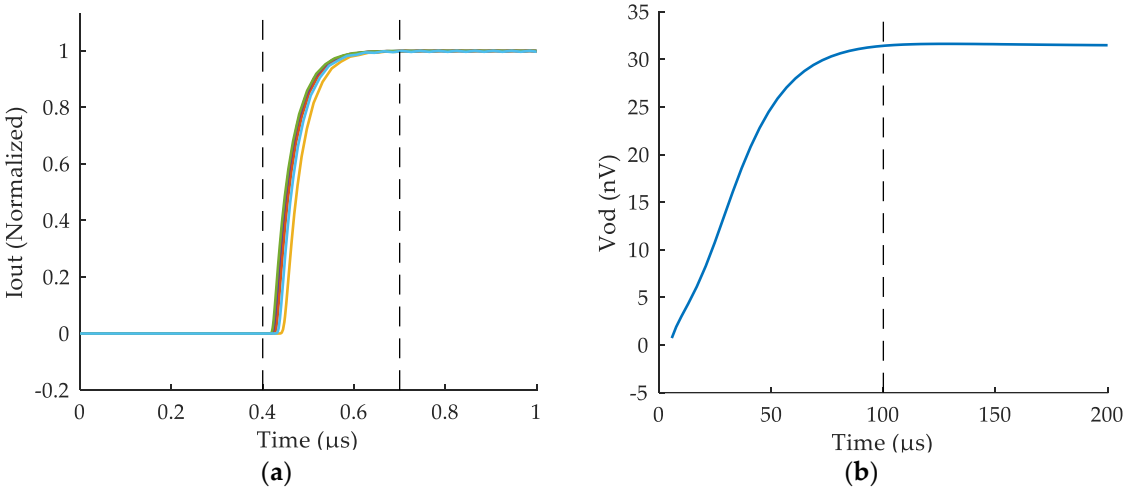

Figure 9. (**a**) Settling time of Gaussian convolution units; (**b**) Settling time of MLP.

5.2.2. Energy Efficiency

The energy consumption of the pre-processing circuit is related to the scale and number of Gaussian convolution unit. Here, we take MNIST (28 × 28) images as the example, which means that the number of convolution operation is $N_{con} = 28 \times 28 = 784$. Assuming the convolution mask is 5 × 5 in size, every Gaussian unit contains $N_{multi} = 25$ multipliers, which consume $P_0 = 25 \times 1.8\,V \times 1\,\mu A = 45\,\mu W$ in total. One unit works at speed of $V_0 = 3.3\,MOPS/s$ should take $T_0 = N_{pixel} \div V_0 = 237.58\,\mu s$ to finish the convolution. Thus, the total energy consumption is $E_{total} = P_0 \times T_0 = 10.69\,nJ$ per image and the energy efficiency is $N_{con} \times N_{multi} \div E_{total} = 1.83\,TOPS/J$.

Table 2. Energy consumption of each part in MLP.

Unit	Cell Number	Energy Consumption (μJ)
Multipliers	39,760	9.23
Non-linear activation	50	1.65×10^{-3}
Current mirror	3950	5.93
Total	-	15.16

As concluded in Table 2, image classification circuit mainly contains multiplication, non-linear activation and current mirror units that respectively consume the energy of 9.23 μJ, 1.65 nJ and 5.93

μJ. Total energy consumption is 15.16 μJ, with the operation number of 7.96×10^4 that including multiplications, additions, etc. Thus, the energy efficiency of the flexible MLP circuit is 5.25G OPS/J.

5.3. Fault Tolerance Analysis

To evaluate the effect of the computational error of the circuit, DoG keypoints matching the experiment is performed between the analog circuit output and software results. As described in the Section 2.2.1, DoG is an approximation of Laplace of Gaussian (LoG) operator, which is usually used to detect the edge of objects in image with the advantages of high efficiency and low complexity. In these experiments, the two filtered images that are obtained by either circuit or the software are subtracted to get the DoG image, see Formula (11):

$$D(x, y) = I_{\sigma 2}(x, y) - I_{\sigma 1}(x, y) \quad (11)$$

where D is the DoG image, and I is the filtered image in the corresponding scale coordinate. The local extremums are detected in 3 × 3 window. If two extremums appear in the same position in the circuit result and the software result, we say they are matched. The matching result is shown in Figure 10 and Table 3.

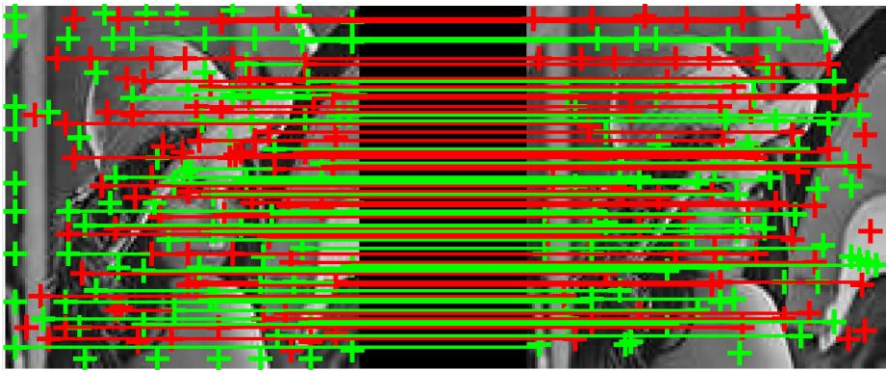

Figure 10. Matching of circuit and software output. The red plus signs are the local maximum points and green ones are the minimum points. The lines connect the matched points in the circuit and software output images.

We can see that the fault caused by analog circuit indeed has the impact on the image feature, especially on the minimum distribution. The interesting thing is that the mismatch mainly happens on the edge of image, according to Figure 10, which is because the large error appears on edge pixels due to the edge information missing we stated in the previous section. But, in some applications, the 76.2% overall matching rate is quite enough to detect the object in the vision. This, in fact, utilizes the inherent fault-tolerance characteristic of the image processing applications. We believe that with the development of Oxide TFT technology, and more elaborate circuit design in future, the computational accuracy could be improved.

Table 3. Matching experiment result.

Key Point	Circuit	Software	Matching	Matching Rate
Maximum	70	60	50	83.3%
Minimum	82	62	43	69.4%
Avg. Rate	-	-	-	76.2%

Mismatch of transistors in the differential pair that introduced by unstable Oxide TFT would result in the nonlinearity of multiplying units and the biasing of the sigmoid function. In order to verify the fault-tolerance of proposed neural network that is based on the flexible device, classification accuracies with different degrees of mismatch are simulated and shown in Figure 11, where the mismatch is set to a uniform distribution of different thresholds. It is worth noting that

the mismatch verification is sufficient to simulate the system function in real applications. Because the system error is mainly introduced by the variations of multipliers and non-linear activation units, of which the differential pair structure makes the variations could be equaled to the mismatch of the transistor pair. That is, the non-ideal factors, such as process variation, temperature, bending, etc., which introduces the error of computation units, can be equivalent to the mismatch of pair transistor to simulate the real applications.

As the maximum threshold of mismatch increases from 1% to 10%, initial error increases as well. Nevertheless, error rates of the mismatch under 5% eventually converge to close to 9%, which is only 2% lower than the software implementation at the same network size. The system still has fast convergent trend within the mismatch of 10%. Therefore, the training of neural network has enough capability to correct the error introduced in the calculation process.

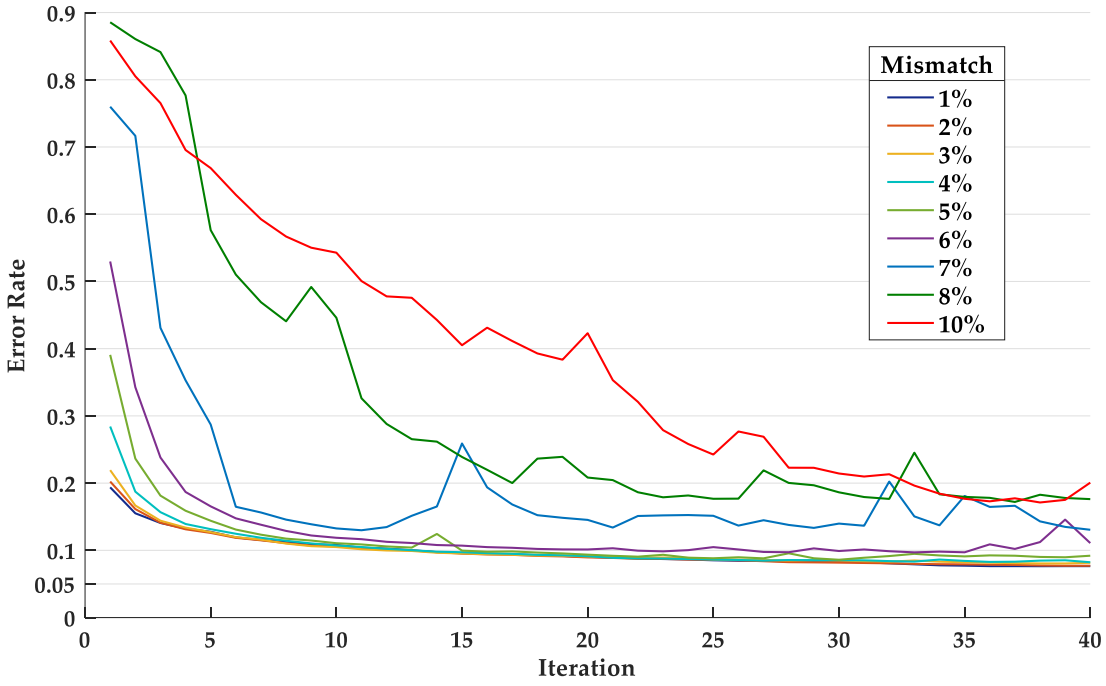

Figure 11. Error rates under different mismatch.

6. Conclusions

In this paper, the problem that flexible electronics are not suitable for computation has been solved. We have explored the computable circuit based on the relatively slower flexible Oxide TFT devices by analog accelerators, which include image filtering and features processing. It is generally considered that analog design is not suitable for flexible devices due to its poor carrier mobility and stability, but not necessarily for some inherently fault-tolerance applications, such as DoG and MLP. In this work, we construct analog circuit unit to realize real-time and power-efficient pre-processing and image classification. Simulation experiments based on Oxide TFT model have demonstrated that the flexible system could perform 5 × 5 Gaussian convolution operation at a speed of 3.3 MOPS/s with the energy efficiency of 1.83 TOPS/J and realize image classification at a speed of 10 k fps, with the energy efficiency of 5.25 GOPS/J. The simulation results show the robustness of the system to imprecise analog domain processing. This work would lead to a new application era of flexible devices to make low power and low cost full flexible systems with sensors, drivers, data converters, and more importantly, the flexible real-time computing parts.

Acknowledgments: The authors would like to acknowledge support from National Natural Science Foundation of China under grant No. 91648116. The authors are grateful to Yingying Zhuang from International school, Beijing University of Posts and Telecommunications for assistance in article structure and investigation of literature.

Author Contributions: Qin Li and Zheyu Liu contributed equally to this work. Qin Li wrote the main manuscript text, did the simulation experiment and analyzed the data of feature-processing part. Zheyu Liu did the simulation experiment and analyzed the data of pre-processing part. The project was guided by Fei Qiao, Qi Wei and Huazhong Yang, who also revised the paper and improved the final document.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, X.; Liu, Z.; Zhang, T. Flexible sensing electronics for wearable/attachable health monitoring. *Small* **2017**, *13*, 1602790.
2. Park, J.-S.; Kim, T.-W.; Stryakhilev, D.; Lee, J.-S.; An, S.-G.; Pyo, Y.-S.; Lee, D.-B.; Mo, Y.G.; Jin, D.-U.; Chung, H.K. Flexible full color organic light-emitting diode display on polyimide plastic substrate driven by amorphous indium gallium zinc oxide thin-film transistors. *Appl. Phys. Lett.* **2009**, *95*, 013503.
3. Schwartz, G.; Tee, B.C.K.; Mei, J.; Appleton, A.L.; Kim, D.H.; Wang, H.; Bao, Z. Flexible polymer transistors with high pressure sensitivity for application in electronic skin and health monitoring. *Nat. Commun.* **2013**, *4*, 1859.
4. Green, M.A.; Emery, K.; Hishikawa, Y.; Warta, W.; Dunlop, E.D. Solar cell efficiency tables (version 47). *Prog. Photovolt. Res. Appl.* **2016**, *24*, 3–11.
5. Zhang, Y.; Fu, H.; Xu, S.; Fan, J.A.; Hwang, K.-C.; Jiang, J.; Rogers, J.A.; Huang, Y. A hierarchical computational model for stretchable interconnects with fractal-inspired designs. *J. Mech. Phys. Solids* **2014**, *72*, 115–130.
6. Sun, W.; Zhao, Q.; Qiao, F.; Liu, Y.; Yang, H.; Guo, X.; Zhou, L.; Wang, L. An 8b 0.8 kS/s configurable VCO-based ADC using oxide TFTs with Inkjet printing interconnection. In Proceedings of the IEEE International Symposium on Circuits and Systems, Baltimore, MD, USA, 28–31 May 2017; pp. 1–4.
7. Song, I.; Kim, S.; Yin, H.; Kim, C.J.; Park, J.; Kim, S.; Choi, H.S.; Lee, E.; Park, Y. Short channel characteristics of gallium–indium–zinc–oxide thin film transistors for three-dimensional stacking memory. *IEEE Electron Device Lett.* **2008**, *29*, 549–552.
8. Yoon, J.; Lee, S.M.; Kang, D.; Meitl, M.A.; Bower, C.A.; Rogers, J. Heterogeneously integrated optoelectronic devices enabled by micro-transfer printing. *Adv. Opt. Mater.* **2015**, *3*, 1313–1335.
9. Myny, K.; Smout, S.; Rockele, M.; Bhoolokam, A.; Ke, T.H.; Steudel, S.; Obata, K.; Marinkovic, M.; Pham, D.-V.; Gulati, A.; et al. 30.1 8b Thin-film microprocessor using a hybrid oxide-organic complementary technology with inkjet-printed P 2 ROM memory. In Proceedings of the 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, USA, 9–13 February 2014; pp. 486–487.
10. LiKamWa, R.; Hou, Y.; Gao, J.; Polansky, M.; Zhong, L. RedEye: Analog ConvNet image sensor architecture for continuous mobile vision. In Proceedings of the 43rd International Symposium on Computer Architecture, Seoul, Korea, 18–22 June 2016; pp. 255–266.
11. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, p. 1150.
12. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
13. Verhelst, M.; Bahai, A. Where Analog Meets Digital: Analog-to-Information Conversion and Beyond. *IEEE Solid-State Circuits Mag.* **2015**, *7*, doi:10.1109/MSSC.2015.2442394.
14. Wu, N.; Liu, Z.; Qiao, F.; Wei, Q.; Guo, X.; Xie, Y.; Yang, H. A Real-Time and Energy-Efficient Implementation of Difference-of-Gaussian with Flexible Thin-Film Transistors. In Proceedings of the 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Pittsburgh, PA, USA, 11–13 July 2016; pp. 455–460.
15. Liu, Z.; Wu, N.; Qiao, F.; Wei, Q.; Guo, X.; Liu Y.; Yang, H. Computable flexible electronics: circuits exploring for image filtering accelerator with OTFT. In Proceedings of the 2016 7th International Conference on Computer Aided Design for Thin-Film Transistor Technologies (CAD-TFT 2016), Beijing, China, 26–28 October 2016; p. 1.

16. Li, Q.; Liu, Z.; Qiao, F.; Wu, X.; Wang, C.; Wei, Q.; Yang, H. From "MISSION: IMPOSSIBLE" to mission possible: Fully flexible intelligent contact lens for image classification with analog-to-information processing. In Proceedings of the 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, USA, 28–31 May 2017; pp. 1–4.
17. Xie, H.; Liu, G.; Zhang, L.; Zhou, Y.; Dong, C. Amorphous Oxide Thin Film Transistors with Nitrogen-Doped Hetero-Structure Channel Layers. *Appl. Sci.* **2017**, *7*, 1099.
18. Nomura, K.; Ohta, H.; Takagi, A.; Kamiya, T.; Hirano, M.; Hosono, H. Room-temperature fabrication of transparent flexible thin-film transistors using amorphous oxide semiconductors. *Nature* **2004**, *432*, 488–492.
19. Bae, J.U.; Baeck, J.H.; Yun, P.; Kim, D.H.; Jang, Y.H.; Park, K.S.; Yoon, S.Y.; Kang, I.B. High mobility oxide TFT for OLED pixel circuits. In Proceedings of the 2017 24th International Workshop on Active-Matrix Flatpanel Displays and Devices (AM-FPD), Kyoto, Japan, 4–7 July 2017; pp. 309–311.
20. Rumelhart, D.E.; Mcclelland, J.L. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations (Parallel Distributed Processing)*; MIT Press: Cambridge, MA, USA, 1986.
21. Lindeberg, T. Scale-space for discrete signals. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 234–254.
22. Kang, K.; Shibata, T. An on-chip-trainable Gaussian-kernel analog support vector machine. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2010**, *57*, 1513–1524.
23. Chandramoorthy, N.; Swaminathan, K.; Cotter, M.; Li, X.; Narayanan, V.; Palit, I.; Irick, K. Understanding the landscape of accelerators for vision. In Proceedings of the 2014 IEEE Workshop on Signal Processing Systems (SiPS), Belfast, UK, 20–22 October 2017; pp. 1–6.
24. Li, Y.; Qiao, F.; Wei, Q.; Yang, H. Physical computing circuit with no clock to establish Gaussian pyramid of SIFT algorithm. In Proceedings of the 2015 IEEE International Symposium on Circuits and Systems (ISCAS) 2015, Lisbon, Portugal, 24–27 May 2015; pp. 2057–2060.
25. Gilbert, B. A precise four-quadrant multiplier with subnanosecond response. *IEEE J. Solid-State Circuits* **1968**, *3*, 365–373.