VNU Journal of Science:
Computer Science and Communication Engineering
Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

On-Chip All-Optical Haar Transform Based on A 4x4 MMI Coupler Cascaded with a 2x2 MMI Coupler for Image Compression

Bui Thi Thuy[1], Dang The Ngoc[2], Le Trung Thanh[3,*]

[1]*FPT University, High-Tech Research Park, Thach That, Hanoi, Vietnam*
[2]*Posts and Telecommunications Institute of Technology (PTIT), Km 10, Nguyen Trai, Hanoi, Vietnam*
[3]*International School (VNU-IS), Vietnam National University, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

Received 05 August 2022
Revised 23 August 2022; Accepted 23 August 2022

Abstract: We present a new method for image compression in all-optical domain. The new hardware architecture is suitable for directly integrating with digital cameras for image processing. The proposed architecture is based on the optical Haar wavelet transform (HWT) using only one 4x4 multimode interference (MMI) coupler cascaded with a 2x2 MMI coupler. The proposed structure can provide a large fabrication tolerance of ±2 μm in length, which is suitable for the current existing CMOS (Complementary Metal Oxide Semiconductor) technology. The numerical simulations show that the compression algorithms have been implemented successfully in all-optical domain. The compression ratios from 10-50% are presented with the Peak Signal to Noise Ratios of 40-131 dB. The processing of images therefore is at very high speed.

Keywords: Image Compression, Haar Wavelet Transform, Signal Transform, Optical Signal Processing, Optical Image Processing.

1. Introduction

Data compression, which is a practical method that helps reduce the use of many expensive resources, is the process of encoding information using fewer bits than the original version. In recent years, the need for efficient compression has increased in line with the requirement for visually beautiful, engaging user platforms with rapid response times to reduce user waiting.

* Corresponding author.
E-mail address: thanh.le@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.446

The implementation of artificial intelligence (AI) algorithms, particularly deep ANN models, still relies heavily on electronic computing now. Even though the individual hardware architectures vary, they always use sophisticated logic circuits and processor chips to complete computations according to the von Neumann computing paradigm. The original neural network architecture relied on a CPU for computation, but it was unable to handle the demands of deep networks' extensive usage of floating-point operations, particularly during training. Moreover, the parallel computing efficiency was too low, and it was quickly replaced by GPU with strong parallel computing capability. It can be said that GPU promoted the development of deep learning. However, the demand for computational power in deep learning is endless. Limiting by the interference of electrical signals, energy consumption and physical limits, although electronic components base on silicon can still support it now, the traditional deep learning has quietly appeared a bottleneck. The academia and industrial circles attempt to seek alternative methods to solve electronic defects that can take precautions on computing power.

The post-law Moore's age, where processing capabilities are no longer advancing as they did over the past several decades, is where electronic accelerator architectures are starting to encounter fundamental limits. A significant bandwidth and energy constraint in these accelerators is the electronic transmission of data across metallic cables. The von Neumann design, which separates memory from processing parts, has the intrinsic drawback of requiring huge amounts of data to be moved, along with the resulting bandwidth and energy needs. One of the most promising answers to solve problems with data transportation is photonic interconnects [1]. At practically every level of the computing hierarchy, photonic linkages have already supplanted metallic ones for the transfer of information at light speed, and they are currently being investigated for integration at the chip size.

With the increased demand for imaging speed, the capture, storage, and processing of image data in the electronic domain presents a serious bottleneck [2]. By transferring some of these conventional signal processing tasks such as buffering, digitization, transformation, and data compression to the photonic domain it is possible to achieve a significant reduction in the electronic workload. In particular, real-time linear transforms, as one of the most fundamental signal processing tasks, take a significant amount of processing power on CPUs and FPGAs. This limited electronic throughput has motivated many groups to investigate electro-optic alternatives [3].

In order to keep video and image processing in the optical domain as much as it is possible, improved techniques and processors are being designed, exploiting optical integrated devices and processing techniques [4]. The main achievable advantages are the high speed, the low losses, the use of guided signals and the possibility to implement real time functionalities [5]. Therefore, a whole new approach of computing has been developed, in which photons are employed as the information carrier rather than electrons. Since the development of laser, the possibility that optics may replace electronics in digital image processing and computation has been investigated.

In reality, photonic computing refers to the manipulation of discrete numerical data. Compared to all-electronic computing, this presents a significant potential for multiple-order-of-magnitude performance increases in areas like speed, non-interfering connectivity, huge parallelism, dependability, and two-dimensional data representation.

The primary appeal of photonic computing is its capacity to represent two-dimensional arrays of discrete binary data using the on/off states of light sources, which, accordingly, represent binary values of one or zero. For a range of computationally intensive issues, like as signal and image processing, equation solving, numerical processing, and the development of digital logics, among other elds, the speed and

parallelism inherent in photonic systems have proven beneficial. The two-dimensionality of photonic systems, the speed of photonic devices, and interconnects are predicted upon by photonic digital computing and processing, which proposes the new and radically different computer architecture.

The architectural concepts include everything from implemented programmable photonic logic arrays using smart pixels, and special purpose cellular image processing, to digital photonic computers based on non-linear logic with new types of optical logic gates.

Image capture, processing, and characterization at high speeds have revolutionized industries such as high throughput microscopy and machine vision. Traditionally, image capture is conducted using CMOS image sensors or CCDs in the electrical domain [6]. Unfortunately, these gadgets have two significant drawbacks: Due to the sluggish rate of electronic data transmission, the frame rate of array-based detectors is restricted to a few megahertz of continuous reading. Second, pixel exposure duration is dependent on device charging time and cannot be shortened arbitrarily, resulting in blurred images and motion artifacts. Utilizing fiber-optic technology, recent research has focused on addressing these weaknesses. However, these systems generate a large quantity of data that must be processed and stored, and although photonic time stretch is a very effective tool for picture acquisition, it does not take full use of the signal processing capabilities provided by photonic technologies in the analog domain.

Image compression is an essential aspect of image processing that may be accomplished via discrete transforms, namely the Haar transform. With the premise that accuracy loss is acceptable, an image compressor is a vital technique that may significantly aid in reducing file size and bandwidth use. The sizes of the arrays are specified as powers of two. The original resolution of the photos is mathematically transformed to the next greater power of two, and the array sizes are initialized correspondingly. The Haar transform splits an image into components of high frequency and low frequency.

Lossless and lossy picture compression may be accomplished efficiently using Haar wavelet compression. It depends on averaging and differentiating image matrix values to build a sparse or nearly sparse matrix. A sparse matrix is one in which the majority of its elements are zero. A sparse matrix may be efficiently stored, resulting in decreased file sizes.This study present a new method based on only one 4x4 MMI coupler cascaded with a 2x2 MMI coupler for image compression directly in all-optical domain for the first time. The architecture can be implanted in the AI camera for image processing before transmiting to the other networks. We design the hardware architecture on the Si3N4 material that suitable for RGB color of images. Therefore, if image processing algorithms can be implemented directly in optical domain, such methods can improve for optical neural networks for image classification and pattern algorithms.

The integration of a complex optical system into a monolithic substrate brings many advantages. With the reduction of size and volume, the stability achieved by electronically controlling many degrees of freedom, and the possibilities to scale up manufacturing in high-yield lithography processes, many optical systems that were previously secluded to research laboratories, can now be ported to industrial and commercial applications. Our proposed method has advantages of low loss, compact size, low fabrication tolerance and high bandwidth.

2. Theory of Image Compression Based on the Haar Transform (HT)

The image array is splitted into two halves containing the transformed data and the detail coefficients. The transformed data coefficients are the results of the low-pass filter while the detail coefficients are the results of the high-pass filter. After transforming the image in the row,

the image is then transformed along the column. Figure 1 shows the working principle of image compression based on discrete Haar wavelet transform (HT).

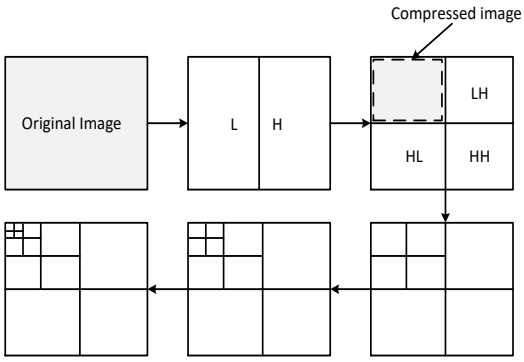


Figure 1. Principle of image compression based on Haar wavelet transform.

An architecture for all-optical image acquisition, processing and transmitting, whose building blocks are depicted in Figure 2 [5]. This approach allows handling images with fast signal processing, maintaining all functionalities in the optical domain.

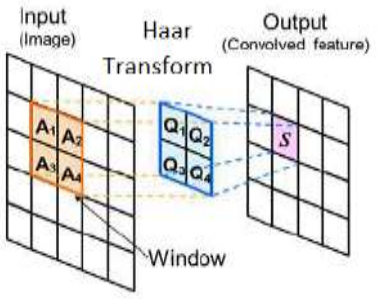


Figure 2. Image compression based on Haar wavelet transform implemented in optical domain.

In recent years, we have presented an approach to implement all-optical Haar wavelet transform using multimode interference (MMI) structure for the first time [7]. From this idea, the Haar transform using directional couplers in all-optical domain for image compression has been developed in recent years [6].

Here, in this study we further develop the Haar wavelet transform for image processing applications. The silicon nitride Si3N4 working at the red, blue and green wavelengths (532 nm, 635 nm and 405 nm) is used for the design. Our structure can be useful for high speed image processing and compression of big data. The new proposed method has advantages of high speed, low loss and compatible with CMOS technology. The Si3N4 platform can work over a long range of wavelength from visible range for human to optical communication systems [8]. The refractive index of Si3N4 is high (2.16), so the compact size and low loss can be achieved.

The Haar transform (HT) is a discrete wavelet transform. HT is a mathematical process using the Haar wavelet. The study of HT will give a solid foundation for discovering more complicated wavelet transformations. As stated before, HT may be utilized for data compression by ignoring less relevant transform domain components. In addition to HT, various more powerful wavelet transformations may be employed for data compression, although it should be noted that HT is simpler to implement due to its simplicity. One of the distinguishing characteristics of HT is its usability for completing basic hand computations [9]. This section describes HT for discrete signals and the signal compression based on it. The HT technique may be used to effectively compress digital photos. Since the picture consists of two-dimensional data, HT must be applied in both dimensions. For instance, for a pixel intensity matrix of a grayscale picture, HT should be performed first to the matrix rows and then to the resultant matrix's columns. In this way, the two-dimensional HT of the image is obtained [9].

As all wavelet transforms, HT decomposes a discrete signal into two sub-signals with a length equal to half the original signal length. One sub-signal is a running average or trend, and the other one is a running difference or fluctuation. The Haar transform is useful in applications where real time implementation of edge detection or contour extraction is required [10]. We now propose a synthesis method for the realization of Haar transforms. This method is suitable for pipeline implementation.

The continuous wavelet transform for a continuois signal f(t) is given by [11]

$$CWT_f(\tau,a)=\frac{1}{\sqrt{a}}\int f(t)h^*(\frac{\tau-a}{a})dt \quad (1)$$

where $h(t-\tau)/a$ is a the mother wavelet shifted and scaled by a variable τ and a, respectively. To discreatize the wavelet transform i a binary, shift and scale varibales defined as $a=2^j$, the mother wavelet is given by:

$$h_{j,k}=2^{-j/2}h(2^{-j}t-k) \quad (2)$$

The DWT for a discrete time signal f(n) can be expressed by

$$DWT_f(j,k)=\sum_n f(n)h_{j,k}(n) \quad (3)$$

The Haar transform decompose a discrete time signal into two sub signals of equal length: a running average and difference. For a discrete signal f with the length of N, the first running aversage and the difference are given by

$$a^1=(a_1,a_2,...,a_{N/2}) \quad (4)$$

$$d^1=(d_1,d_2,...,d_{N/2}) \quad (5)$$

$a_m=\frac{f_{2m-1}+f_{2m}}{\sqrt{2}}$, $d_m=\frac{f_{2m-1}-f_{2m}}{\sqrt{2}}$, for m=1, 2, 3,...,N/2

The first order of the Haar transform is the mapping $f^{H_1}\mapsto H_1=(a^1,d^1)$. The Haar transform is based on the Haar function, which is periodic, orthogonal and complete. The first order Haar matrix is defined as

$$H_1=\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix} \quad (6)$$

3. All-optical Haar Transform Based on 4x4 MMI and 2x2 MMI Couplers

As mentioned earlier, optical HT can be realized using MMI based structures which have some advantages such as high bandwidth, polarization insensitivity, compact size and convenient integration, easy fabrication and low optical loss [7]. Different type of MMI couplers and their design relations have been presented in detail in [12]. The multimode interference coupler consists of three parts: a single mode input, output waveguides, and a multimode waveguide linking the two input and output regions. The multimode area has a broad breadth to accommodate many modes. The MMI coupler operates based on the Talbot effect. There are three basic interference processes based on the positions of input field excitations. The primary mechanism is general interference (GI), which is independent of modal excitation. The second is the limited interference (RI) mechanism, in which excitation inputs are positioned in certain locations to prevent the stimulation of particular modes. The last mechanism is symmetric interference (SI), in which the excitation input is centered inside the multimode region.

In this study, the access waveguides are identical single mode waveguides. The input and output waveguides are located at [13]

$$x_i=(i+1/2)W_{MMI}/N \quad (7)$$

We have showed that the characteristics of an MMI device can be described by a transfer matrix matrix . This transfer matrix is a very useful tool for analysing cascaded MMI structures. In this study, we propose a 4x4 MMI coupler with a width of W_{MMI}, length of $L_{MMI}=1.5L_\pi$ to realize the expected matrix of the MMI. The transfer matrix of the 4x4 MMI coupler is

$$M=\begin{bmatrix}1-j & 0 & 0 & 1+j\\ 0 & 1-j & 1+j & 0\\ 0 & 1+j & 1-j & 0\\ 1+j & 0 & 0 & 1-j\end{bmatrix} \quad (8)$$

As an example, the second Haar matrix is given below

$$H_2=\begin{bmatrix}1 & 1 & 1 & 0\\ 1 & 1 & -1 & 0\\ 1 & -1 & 1 & 1\\ 1 & -1 & 0 & -1\end{bmatrix} \quad (9)$$

This matrix can be obtained from the 4x4 and 2x2 cascaded matrices. The matrix of the 4x4 MMI coupler can be rewritten by

$$S = \frac{1}{\sqrt{2}} e^{\frac{-j\pi}{4}} \begin{bmatrix} 1 & 0 & 0 & \exp(j\frac{\pi}{2}) \\ 0 & 1 & \exp(j\frac{\pi}{2}) & 0 \\ 0 & \exp(j\frac{\pi}{2}) & 1 & 0 \\ \exp(j\frac{\pi}{2}) & 0 & 0 & 1 \end{bmatrix} \quad (10)$$

As a result, the 4-point Haar as shown in Figure 3. In Figure 3, the structure requires transform can be realized by cascading three stages of 2-point Haar transform three 2-point Haar transforms and an exchange unit. In this study, we present a new method based on cascaded 4x4 MMI and 2x2 MMI, which does not require cross-over waveguide (Figure 4).

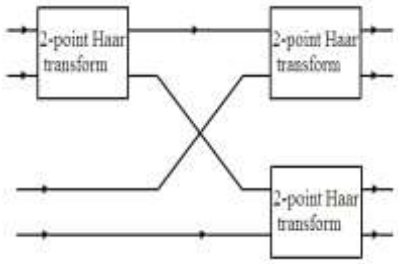


Figure 3. 4x4 Haar transform using 2-point Haar transforms.

In this study, we propose a new 4-point Haar transform using only 4x4 MMI cascaded with a 2x2 MMI as shown in Figure 4.

Figure 4. Photonic circuit for implementing the 4-point Haar transform using 4x4 MMI cascaded with a 2x2 MMI coupler.

4. Simulation Results and Discussions

In this study, the Si3N4 waveguide working at visible wavelengths RGB colour is used. The height and width of the waveguide are 170 nm and 1600 nm, respectively (Figure 5). At the visible wavelength, we choose these height and width of the waveguide for single mode operation. For multimode waveguide, we use wider width. The calculated effective refractive index of the single mode waveguide using the BPM (Beam Propagation Method) simulation is to be Neff=1.7.

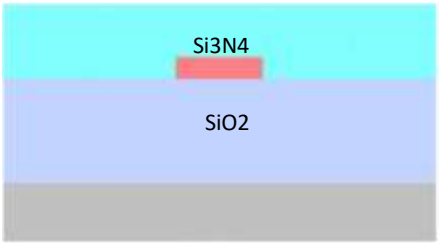
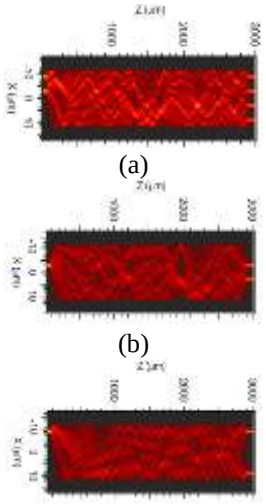


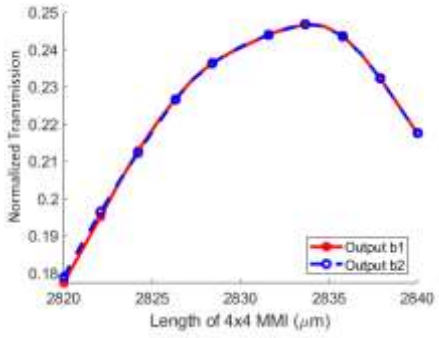
Figure 5. Waveguide structure.

The numerical simulation results for the 4x4 MMI design are shown in Figure 6. The optimal length and width of the 4x4 MMI calculated to be 2833 μm and 24 μm. Figure 6a shows for input signals at port 1 and 2. Figure 6b and 6c shows the input signal at port 1 or port 2, respectively.

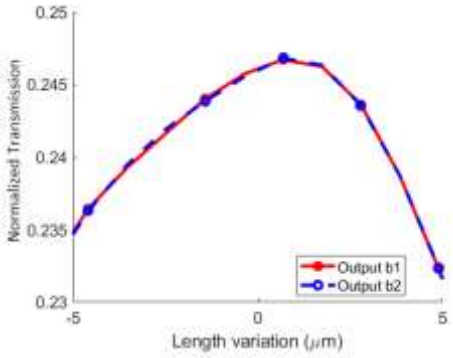


(a) (b) (c)

Figure 6. (a) Input signal at port 1, 2, (b) input signal at port 1 and (c) input signal at port 2.



(a)



(b)

Figure 7. (a) Powers at port 1 and 2 when input signal is at port 1 and port 2 with different lengths of the 4x4 MMI and (b) Fabrication tolerance analysis for variation of the 4x4 MMI length.

The normalized powers at output ports 1 and 2 when input signal is at port 1 and port 2 is shown in Figure 7. The simulations show that the length variation of ±2 µm is still keep the output powers unchanged. This means that the fabrication tolerance of the proposed structure is high. The current CMOS fabrication technology for VLSI industry is feasible.

Next, we investigate the phase error of the Haar transform. The phases at output ports 1 and 4 when input signal is at port 1 are shown in Figure 8. The phase shift difference between port 1 and 4 is also presented in this simulation. The results show that the phase difference of 90 degree can be obtained over a length variation of 18 µm. This result provide a flexible design for the optical HT. The HT can be implemented extremely accurately using the CMOS fabrication technology.

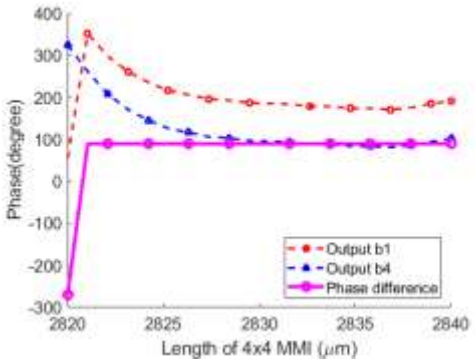
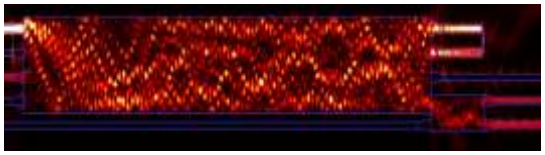
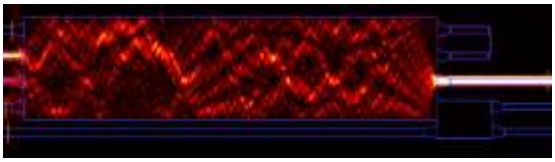


Figure 8. The phases at output ports 1 and 4 when input signal is at port 1.

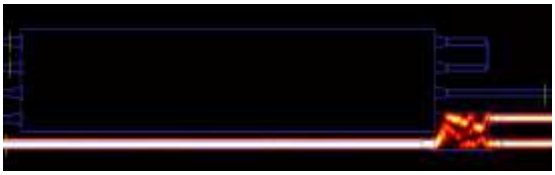


(a) input signals 1000

(b) input signals 0100



(c) input signals 0010

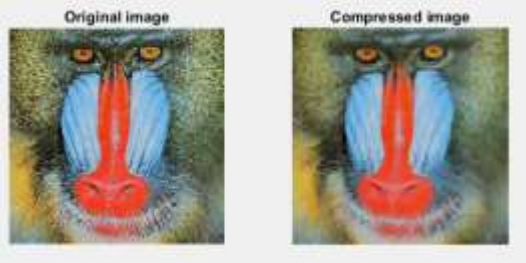


(d) input signals 0001

Figure 9. The 3D simulations for the 4-point Haar transform with different input signals (a) input signals “1-0-0-0”, (b) input signals “0-1-0-0”, (c) input signals “0-0-1-0” and (d) input signals “0-0-0-1”.

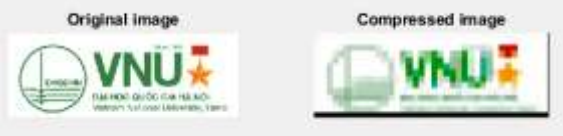
We need to find out the optimal length and width of the MMI for image processing. The 3D

simulations to find the optimal length of the 4-point Haar transform using the 4x4 coupler and 2x2 MMI coupler are shown in Figure 8a and in Figure 8b, respectively.



(a) "Mandril" image, CR=10%

(b) "Lena" image, CR=21%



(c) "VNU Logo" image, CR=50%

Figure 10. Original and compressed images.

The image processing going over the MMI is simulated in Figure 9. The first case is for input light beams applied to input port 1. The 3D-BPM simulation for this case is plotted in Figure 9a. The second case is for an input signal presented at input port 2 and the third case is for input signals presented at input ports 1 and 2. The simulations for these cases are presented in Figure 9b and 9c. The simulations show that the device performs the functions of the 4-point Haar transform as predicted by the theory. The excess loss of this device is 0.95 dB for the first case, 0.78 dB for the second case and 0.48 dB for the third case.

As an example, we show the results for compressed images compared to the original image with different compressed ratios (CR) in Figure 10. Performance evaluation is based on the Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR) calculation. The MSE between two images f and g (size MxN) can be expressed by

$$MSE = \frac{1}{MxN}\sum_{j=1}^{N}\sum_{k=1}^{M}\left[f(j,k)-g(j,k)\right]^2 \quad (11)$$

$$PSNR\ [dB] = 10Log_{10}\left(\frac{255^2}{MSE}\right) \quad (12)$$

Table 1 presents the comparison of the MSE, PSNR, file size for different input images compressed by using the optical Haar transform. The MSE will increase when the compression ratio is increased, while the PSNR will decrease. These simulations are suitable with equations (11) and (23) due to the loss compression methods based on the Haar transform.

Table 1. Image compression results for different images

Images	Original Size	Comp. size	CR	MSE	PSNR (dB)
Mandril	787kB	785kB	10%	0.23	140
Lena	787kB	378kB	21%	0.37	131
VNU Logo	19kB	10kB	50%	34	40

5. Conclusions

This paper have presented an new approach for implementing image compression technique based on Haar wavelet transform using the 4x4 MMI coupler cascaded with a 2x2 MMI coupler in all-optical domain. The proposed approach is

useful for image processing and big data processing at extreme high speed. The proposed method can be integrated with the AI camera to implement image processing directly in the camera.

Acknowledgments

This research is funded by National Science and Technology Program (Program 562) under grant number ĐTĐLCN-92/21.

References

[1] F. P. Sunny, E. Taheri, M. Nikdast, S. Pasricha, A Survey on Silicon Photonics for Deep Learning, J. Emerg. Technol. Comput. Syst., Vol. 17, No. 4, 2021, pp. 61, https://doi.org/110.1145/3459009.

[2] L. Deligiannidis, H. Arabnia, Emerging Trends in Image Processing, Computer Vision and Pattern Recognition, Morgan Kaufmann, 2014.

[3] M. Papaioannou, E. Plum, N. I. Zheludev, All-Optical Pattern Recognition and Image Processing on a Metamaterial Beam Splitter, ACS Photonics, Vol. 4, No. 2, 2017, pp. 217-222, https://doi.org/110.1021/acsphotonics.6b00921.

[4] A. A. Fashi, M. H. Vadjed Samiei, A. Teixeira, Design of a Visible Light Photonic Chip for Haar Transform Based Optical Compression, Optik, Vol. 217, 2020, pp. 164929, https://doi.org/10.1016/j.ijleo.2020.164929.

[5] L. Almeida, N. Kumar, G. Parca, A. Tavares, A. Lopes, A. Teixeira, All-Optical Image Processing Based on Integrated Optics, 16th International Conference on Transparent Optical Networks (ICTON), https://doi.org/110.1109/ICTON.2014.6876660.

[6] A. A. Fashi, M. H. V. Samiei, C. Pinho, A. L. Teixeira, Photonic Integrated Chip on TriPleX Platform for Realizing Optical Haar Transform and Compression in the Visible Spectrum, IEEE Journalof Quantum Electronics, Vol. 57, No. 5, 2021, pp. 1-10, https://doi.org/110.1109/JQE.2021.3102845.

[7] T. T. Le, The Design of Optical Signal Transforms Based on Planar Waveguides on a Silicon on Insulator Platform, International Journal of Engineering and Technology, Vol. 2, No. 3, 2010, pp. 245-251.

[8] Y. D. Yang, Y. Li, Y. Z. Huang, A. W. Poon, Silicon Nitride Three-Mode Division Multiplexing And Wavelength-Division Multiplexing using Asymmetrical Directional Couplers and Microring Resonators, Optics Express, Vol. 22, No. 18, 2014, pp. 22172-22183, https://doi.org/110.1364/oe.22.022172.

[9] A. A. Ensafi, F. Akbarian, E. H. Soureshjani, B. Rezaei, A Novel Aptasensor Based on 3D-Reduced Graphene Oxide Modified Gold Nanoparticles for Determination of Arsenite. Biosensors & Bioelectronics, Vol. 122, pp. 25-31, https://doi.org/10.1016/j.bios.2018.09.034.

[10] S. Burrus, R. A. Gopinath, H. Guo, Introduction to Wavelets and Wavelet Transforms: A Primer, Prentice Hall, United States, 1997.

[11] A. A. Fashi, M. H. V. Samiei, A. Teixeira, Realization of Visible Light Integrated Circuits for All-Optical Haar Transform, Optical and Quantum Electronics, Vol. 53, No. 7, 2021, pp. 364, https://doi.org/110.1007/s11082-021-03008-5.

[12] T. T. Le, Multimode Interference Structures for Photonic Signal Processing, LAP Lambert Academic Publishing, 2010.

[13] J. M. Heaton, R. M. Jenkins, General Matrix Theory of Self-Imaging in Multimode Interference (MMI) Couplers, IEEE Photonics Technology Letters, Vol. 11, No. 2, 1999, pp. 212-214.