<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Inference engine for expert system using optical array logic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Iwata, Masaya; Tanida, Jun; Ichioka, Yoshiki</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Applied Optics. 31(26) P.5604-P.5613</td>
</tr>
<tr>
<td><strong>Issue Date</strong></td>
<td>1992-09-10</td>
</tr>
<tr>
<td><strong>Text Version</strong></td>
<td>publisher</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/11094/3286">http://hdl.handle.net/11094/3286</a></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
<td></td>
</tr>
<tr>
<td><strong>rights</strong></td>
<td></td>
</tr>
</tbody>
</table>

*Osaka University Knowledge Archive : OUKA*

https://ir.library.osaka-u.ac.jp/repo/ouka/all/

Osaka University
Inference engine for expert system by using optical array logic

Masaya Iwata, Jun Tanida, and Yoshiki Ichioka

A method of implementing an inference engine for an expert system by using optical array logic is presented. Optical array logic is a technique for parallel neighborhood operations based on spatial coding and two-dimensional correlation. The processing capability of optical array logic is useful for the efficient execution of an inference operation in artificial intelligence problems. Inference proceeds with the help of a token propagation technique in which the knowledge is expressed as spatial patterns and processed collectively with parallel neighborhood operations in optical array logic. With several additional functions, an expert system can be constructed by using the inference technique.

Key words: Digital optical computing, optical array logic, artificial intelligence, inference engine, expert system.

1. Introduction

Optical parallel computing systems are capable of massive parallel processing and of making good use of the inherent parallelism and the high speed of light propagation. The systems are expected to solve problems in massive information processing. Therefore various studies on system architecture and computational algorithms in optical computing have been reported.1-3

Optical array logic (OAL) is proposed as a technique for achieving a parallel neighborhood operation for two binary images.4,5 The advantages of OAL are its capability of parallel logic operations, its programmability with an operation kernel, and its suitability for optical parallel processing. Several optical systems have been developed to implement OAL.6,7 On these optical systems, various single-instruction stream–multiple-data stream (SIMD) types of parallel processing can be achieved with OAL. Therefore studies of parallel processing with OAL programming are important subjects for exploring the capability of OAL processors. Image and numerical processing have been studied4-8 on the basis of this principle.

Artificial intelligence9 (AI) is one of the research fields in which massive parallel processing is strongly desired. The main process in the AI problem is inference. Inference is an operation that searches specific data from a knowledge base, and usually it must handle a large amount of data. In a practical system, such as an expert system,10 this inference operation determines the total performance of the system. Thus it is important to develop an efficient searching method for AI problems.

Recently several methods have been proposed for a data search that utilize the parallelism in optics.11-15 In some methods, knowledge data are represented in an image, and inference is achieved by SIMD types of parallel operation such as matrix operations12 and matched filtering.13 Since these operations are effectively executed by optical systems, a high throughput of inference can be expected. We also found that OAL is useful for parallel data search in inference. In Ref. 14 we reported an inference engine that uses a template matching technique based on OAL. However, these methods lack flexibility and, hence, cannot achieve minute operations for data. To overcome these shortcomings, there is an approach to map knowledge data directly onto an optoelectronic multiprocessor system.15 In this method, inference is executed by marker propagation among processing elements with an optical interconnection. However, the parallelism of light is not fully utilized in this method.

In this paper we present a method of implementing an inference engine with a token propagation technique that is a useful technique in OAL programming. This method can fully extract parallelism in optical processing and can be executed by an OAL processor. The flexibility of this method is demonstrated by an
expert system that uses the inference engine. In Section 2 we briefly explain OAL and a specific operation used in an inference process. In Section 3 we describe an inference process based on a semantic network as knowledge representation. In Section 4 we explain the concept of token propagation and we discuss a method of implementing a primitive inference engine by using the technique. In Section 5 we describe an inference engine for an expert system, and in Section 6 we estimate the processing efficiency of the presented inference engine.

2. Optical Array Logic and Template Matching for Two Images

OAL is a technique for executing logic operations for a pixel and their neighboring pixels on two-dimensional (2-D) binary images in parallel. The processing procedure of OAL is shown in Fig. 1. Two binary images are inputs on which all pair of pixels at the same location are converted into spatial patterns by the coding rule, which is shown in the Fig. 1. The resultant image, which is called a coded image, is correlated with an operation kernel configuring operation of OAL. The operation kernel is a set of delta functions positioned at some grid points in which the interval of the grid is equal to the cell size of the coded image. Therefore the correlation is regarded as a sequence of operations: duplicating the coded image, and then shifting and overlapping these images.

The correlated image is spatially sampled at the upper left pixels of cells on the input images. As a decoding process, the collection of the sampled pixels is dilated and inverted. By the above sequence, we can execute a product term operation that is a range of logic operations expressed by the logic product of pixel values. To extend the flexibility of the operation, multiple product term operations are executed with different operation kernels, and their results are combined by a logic sum. As a result, OAL is capable of executing an arbitrary neighborhood logic operation with a combination of operation kernels used in correlation. These operations are executed in parallel with a simple optical correlator, as shown in Fig. 2. Thus an SIMD-type of parallel neighborhood operation can be achieved by OAL.

Parallel processing, as described by OAL, can be optically executed on a general-purpose optical processing system called an OPALS (optical parallel array logic system). Figure 3 shows a block diagram of a hybrid version of the OPALS. OAL is considered as software (program) for controlling such an OPALS system. The following is addressed to programming techniques for the OPALS.

Programming OAL is achieved by specifying a set of operation kernels used in correlation. For programming convenience, we have developed kernel expressions that can denote a sequence of operation kernels. A kernel expression is an intuitive expression of logic operations for 2-D images. Thus, if a problem is given as a logic relation between spatial patterns, it can be easily described by a kernel expression. Once a kernel expression is determined, we can convert it into operation kernels used in OAL and execute on the OPALS. The notation of the kernel expression is summarized briefly in appendix A. The operation kernels designed in this study are listed in appendix B.

As an example of OAL, we explain a useful operation in data search for an inference operation. The operation is basically logic template matching, but it operates the patterns in images A and B at the same time. In this operation, specific positions in which the combination of the pixels in two images are identical to the predetermined patterns are detected.

The operation is achieved by a parallel AND operation for neighborhood pixels in the two images. For example, to detect occurrences in which pixels with 1 and 0 are arranged side by side in the corresponding position of images A and B, the following operation is used:

\[ c_{i,j} = a_{i,j} \bar{a}_{i,j+1} b_{i,j} \bar{b}_{i,j+1}, \]

where \( a, b, \) and \( c \) mean pixel data in images A, B, and C, respectively; subscripts denote the location of the data in the image. A kernel expression of Eq. (1) is shown in appendix B. Figure 4 shows the flow of the operation.

In OAL, one input image can be assigned to control signals for the data primitives located in the other
input image. For this technique, template matching on two images is frequently used to select specific data. This template matching plays an important role in the inference operation by OAL.

3. Inference
In a practical AI system, such as an expert system, inference is a process for answering given queries with a knowledge base that consists of data that represents various knowledge. An inference engine is a mechanism for finding a solution for queries by searching data related to the queries with a knowledge base. To execute humanlike intelligent inference, the knowledge base must contain a huge amount of data. Thus searching data in inference must be executed efficiently by fast processing.

There are several methods for knowledge representation in computer science, for example, a production rule, a frame, a semantic network, and so on. We use a semantic network as a knowledge representation because it is suitable for SIMD-type processing in OAL. A semantic network is a diagrammatic knowledge representation that is described by a directed graph consisting of nodes and links [Fig. 5(a)]. A node describes a concept, and a link represents a relation between two concepts. Inference with a semantic network is achieved by tracking links from a node to the connected nodes. For example, when a query "What is a cat?" is given, the answer "An animal" is obtained by reversely tracking the ISA link from the Cat node to the Animal node.

The process of tracking links is executed by a method called marker propagation. A marker is an information primitive that propagates on links to help the inference process. Parallelism in inference with a semantic network is obtained mainly by the marker propagation algorithm.

For example, when a query "What animal has whiskers?" is given, as shown in Fig. 5(b), two markers (marker #1 propagating along the HAS link in the reverse direction and marker #2 propagating along the ISA link in the reverse direction) are prepared and set on Whiskers and Animal nodes, respectively. These markers propagate simultaneously along the specified links. The node to which both markers reach indicates the answer Cat [Fig. 5(c)].

The process of the inference engine with marker propagation is as follows: (1) searching nodes appearing in a given query, (2) setting markers on the searched nodes, (3) propagating the markers along the specified links to the next nodes, and (4) repeating step (3) until all markers come to the end of the propagation paths.

Fortunately the above process can be executed for all markers with a SIMD type of parallel processing. In our method, a token propagation technique is used to implement the markers. The direction and type of links are encoded into pixel patterns on an image plane. As a result the capability of parallel processing in OAL is quite useful for an inference process to use to increase processing efficiency.

---

Fig. 2. Optical correlator that is using a shadow-casting system.

Fig. 3. Block diagram of a hybrid version of the OPALS. SLM, spatial light modulator; LED, light-emitting diode; PD, photodetector.

Fig. 4. Processing flow of template matching for two images.

Fig. 5. Semantic network and inference based on the semantic network: (a) an example of a semantic network, (b) a query, (c) inference based on the semantic network with marker propagation.
4. Primitives for Inference Engine in Optical Array Logic

A. Encoding of a Knowledge Base

To implement an inference process by using OAL, a semantic network must be represented by pixel patterns on an image plane. A semantic network in Fig. 5(a) is converted into a spatial representation by the following procedures, as shown in Fig. 6.

First, the target semantic network is divided into multiple subnetworks with a one-to-many structure. In an individual subnetwork one node is connected to several nodes through different links [Fig. 6(a)]. The assembly of the subnetworks is coded into a table format in which the rows and the columns are assigned to each subnetwork and to an individual node, respectively [Fig. 6(b)].

In the table format, the direction and the type of link must be stored at each cell. The type of link is set at the terminating cell of the link. The direction and type of link are coded into spatial patterns, called node patterns, and set at the cells [Fig. 6(c)]. An area for setting a marker in inference is also prepared to the cell. Code patterns are assigned to each meaning, as shown in Fig. 6(d). In the explanation of the link direction pixels, a small dot indicates a node. The code patterns for the link type are the specified ones for this example. By setting the code patterns to the appropriate position in Fig. 6(b), we can obtain a spatial representation of a semantic network for OAL [Fig. 6(e)].

In addition, to identify the individual cells in the spatial representation, anchor tags are set on the other image [Fig. 6(f)]. This technique is often used in OAL to increase processing flexibility. The images containing the data and the anchor tags are called data and attribute planes of a knowledge base, respectively. Template matching for both planes enables us to access an arbitrary part of the knowledge base.

A query is represented with a subset of a semantic network, as shown in Fig. 5(b). It can be encoded in the same way as the knowledge base. That is, the query is represented by the table format and coded into spatial patterns. Usually each condition in a query is assigned to an individual marker for inference. Thus markers #1 and #2 are set in the appropriate positions in the cells [Fig. 6(g)].

B. Token Propagation

To implement parallel inference by using OAL, we develop a token propagation technique. The basic ideas of the token propagation are (1) setting some tokens on a 2-D image to indicate interested objects and (2) transferring the tokens concurrently to execute desire operations with parallel neighborhood logic operations in OAL. Since individual tokens are transferred independently, great flexibility and processing capability can be attained even on a SIMD processor.

To implement token propagation with OAL, pattern expansion and template matching are employed. Figure 7 indicates a procedure for propagating a token to a desired destination. First, a token to be propagated is duplicated over a specific area. This operation is called pattern expansion and is achieved by the logical sum of variables a and b at the neighborhood location on which the target pattern is to be expanded. Correlation in the procedure of OAL provides this pattern expansion. The expanded tokens are selected by template matching; the template is specified by an image called a condition image. As a result, the initial token is transferred to the destination that satisfies a given condition. This procedure is executed by SIMD operations so that OAL can be used to implement it.

In token propagation, the area of pattern expansion and the condition image are used to specify the contents of the processing. As shown in Fig. 8, if tokens are expanded in the horizontal (or vertical) direction, multiple tokens can be transferred indepen-
Fig. 8. Procedure for propagating multiple tokens to desired destinations.

dently. In addition, appropriate patterns are set on the condition image so that the behavior of each token can be controlled. Although the destination of tokens is limited by the area of pattern expansion, multiple processes are handled in OAL with this technique.

C. Inference with Token Propagation

With the token propagation technique on a spatially encoded knowledge base, inference processing is executed efficiently by OAL. For example, a problem in Fig. 5(b), or a query, “What animal has whiskers?,” is solved by OAL. As described in Section 3, markers propagating from the Whiskers and the Animal nodes are used for this problem. In our method, tokens on an image are assigned to the markers and propagated by OAL operations.

The basic procedure for executing the inference is (1) searching subnetworks corresponding to given queries, which are assigned to the rows in the knowledge base; (2) setting tokens at the searched cells; (3) propagating the tokens in the same rows to track reversely the specified links; and (4) repeating steps (1)–(3) until all tokens reach an identical node or the same column.

Figures 9–11 show practical processing for the inference. Steps (1) and (2) from above are executed by vertical token propagation, shown in Fig. 9, in which the condition image is created from the knowledge base to select the terminating nodes of the ISA and HAS links. In this case, a pattern consisting of $4 \times 2$ pixels for a cell is assigned as a token. To identify the region of the individual cell, anchor tags in the attribute plane of the knowledge base are used. As a preparation of step (3), the link direction pixels on the tokens are changed to indicate an initiating node, as shown in Fig. 10. The modified tokens are propagated horizontally to execute step (3), as shown in Fig. 11. The transferred token is again changed to indicate an initiating node. The obtained token has the same format of the query, so that the above processing is repeated until the final result is obtained, as shown in Fig. 12. The operation kernels used in the operations are shown in Appendix B.
Pattern Expansion

Template Matching

Knowledge Base
Condition Image (2)

Query (2)

Expanded Image (2)

Query (1)

Query (2')

Query (1)

Query (2)

Query (3)

Query (3)

Query (4)

Query (4)

Query (5)

Query (5)

Fig. 11. Processing sequence of horizontal token propagation: (a) images, (b) semantic representation.

Fig. 12. Processing sequence of token propagation in query image: (a) images, (b) semantic representation.

Fig. 13. Implementation of inheritance based on a semantic network: (a) an example of a semantic network, (b) a query, (c) inheritance based on the semantic network with marker propagation.

13(a). As shown in the semantic network, Cat is a member of Animal. Since characteristics of Animal must be held by its member Cat, the property Has Skin should be inherited from the Animal to the Cat node. The function is achieved by propagating markers along both a specified link (in this example, the HAS link) and along the ISA link [Fig. 13(c)].

To execute function (2), markers must be recorded on all the propagated nodes. After propagation, we can find the node that satisfies specific conditions by detecting the corresponding recorded markers. In the example of Fig. 13, by detecting a marker pair of #1 and #2 the Cat node is derived as the answer.

Before detection of the nodes with the desired markers, the marker propagation process must be terminated [function (3)]. This function is achieved by detecting whether any marker is propagated from the current node. Since these functions can be executed for all data at a time, parallelism in OAL is utilized.

B. Inference Engine for Expert System by Using OAL

The processing flow of inference for an expert system is shown in Fig. 14. The main part of the processing is identical to the marker propagation process described in Section 4. Thus we explain additional functions for an expert system.

To simplify the sequence, the inheritance process is assumed to be executed only by tracking reversely. In this case, all links in the given query, except the ISA links, are reversely tracked and then the same operation is executed for the ISA links. As a result, inheritance can be utilized in a kind of inference, such as “What has skin?” Although this restricts the functions of an expert system, a more complicated sequence by OAL enables us to extend the function.

Recording of the propagated markers is achieved by copying the marker number pixels in the query to the corresponding pixels on the knowledge base. Figure 15 shows an example of the knowledge base before
I Input a coded query
Kjmrrprrc
opagation
No
Deetnnos
propagated >awhich have desired
markersd
As
Recording propagated |
&
Imarkers
in the knowledge base|
Inheritance operato
I I
Fig. 14. Processing flow of the inference engine for an expert system with marker propagation.

and after recording markers. This operation can be executed by simple logic operations.

To detect whether any marker is propagated from the current node, we introduce conditional variables as an extension of OAL. Conditional variables indicate conditions of the processed image, like conditional flags in a microprocessor. For this case, a ZERO variable is prepared, whose value is 1 if all pixels in the output image are 0 and is 0 for the other case. Although the same function is implemented by a sequence of logic operations in OAL, conditional variables can greatly improve processing efficiency through the use of specific optical hardware. For example, a ZERO variable is simply implemented with a condenser lens and a photodetector, as shown in Fig. 16. As a result, a conditional test for termination is achieved by referring to the ZERO variable.

The nodes with the desired markers are detected by template matching. In Fig. 15, the marker pair of #1 and #2 is detected by testing the third column of each cell. As a result, the upper left cell is detected, and Cat is derived as the answer.

C. Simulation Result
With the above process, a simple expert system is implemented for a knowledge base, as shown in Fig. 17(a). On the expert system, a query in Fig. 17(b) is asked as an example: “What cat has short hair and no tail?” Figure 18 shows a simulated result of the expert system. Figures 18(a) and 18(b) are the attribute and the data planes of the knowledge base, respectively, and Fig. 18(c) indicates the query image. Figure 18(d) is the data plane of the knowledge base on which markers are recorded after the marker propagation process. Figure 18(e) shows the detected result of the node on which all used markers are recorded. The node is the answer of the query. The vertical and the horizontal scales around the images indicate the borders of the subnetworks and the individual nodes, respectively. The horizontal scale numbers correspond to the identifiers of the

Fig. 16. Optical system for the conditional variable ZERO.

Fig. 17. Semantic network for an expert system: (a) an example of a knowledge base, (b) a query. The numbers in parentheses are the identifiers of the nodes.
6. Performance Estimation

The processing performance of the programs described in OAL can be estimated by three factors: the total number of product terms required in the processing, the spatial size of the operation kernels used in correlation, and the spatial size of the images to be processed. The first determines the system cycles, and the others state necessary hardware specifications for the processing. The total number of product terms corresponds to the total number of system cycles required in processing. Actual processing time is obtained by the product of the total number of product terms and the processing time for one system cycle, which depends on the response time of the devices used, such as spatial light modulators.

Table 1 tabulates these values of the inference engine and the expert system by using the token propagation. As Table 1 shows, these values are greatly affected by the size of the semantic network. For a practical expert system, at least a few hundred nodes are required. If 500 nodes are required, images that consist of 1000 × 1000 pixels and operation kernels with 2000 × 2000 units must be handled in an optical system.

The processing speed is estimated by the total number of product terms in the processing. Note that this value is independent of the size of the semantic network in one cycle of the inference sequence. Consequently, if a large size of images can be processed, the inference engine with OAL has the capability of executing one cycle of the inference sequence in a constant time, regardless of the size of the semantic network.

However, to accomplish an inference process, one step of inference described here must be iterated by the number of the maximum depth of a hierarchy in a semantic network. If the average of the division number of node information is \( m \), the maximum depth of hierarchy is considered to be \( \log_m n \), where \( n \) is the number of nodes. This value depends on the character of a semantic network.

If an image representing a semantic network is larger than that handled by a system, the image must be separated and processed page by page. However, this method needs additional tasks for managing continuity of adjacent pages. Since this process reduces the performance, an efficient processing method of the overlapped pages should be considered.

7. Conclusion

We have presented an implementation method of an inference engine that utilizes the parallelism of OAL. The token propagation technique that was proposed as an OAL programming technique is useful for parallel inference. By using this technique, we have developed OAL programs for an inference engine and an expert system. As a result of the performance estimation, we have found that if a large size of images is handled on an optical system, the inference engine by OAL has the capability of executing one
step of the inference sequence in a constant time regardless of the size of the semantic network.

Appendix A
An operation kernel explicitly corresponds to logical meanings. Thus it can be expressed by symbolic notation for intuitive expression. We call the native notation for OAL kernel expression. The essential components of the kernel expression are as follows:

**Product Term Block: PTB**
Since an operation kernel specifies a product term operation, it is reasonable to introduce a functional block that corresponds to an operation kernel. The functional block is called the product term block, or PTB. It is indicated by a set of brackets.

**Function Symbols**
Elements in a PTB are two-letter symbols that indicate any one of two-variable binary logic functions. The symbols are tabulated in Table 2. The first and second letters assign functions for inputs A and B, respectively. To handle both the logic product and the sum, two series of letter sets are used. 1, 0, and . are for products, whereas P and N are for sums; symbols UU, EE, and DD are special symbols for specific functions.

**Neighborhood Mapping and the Origin Marker**
Operands of the target logic neighborhood operation are specified by the location of function symbols set in a PTB. That is, a PTB directly corresponds to a neighborhood area. The origin of a PTB is indicated by an underscore. If a PTB has only one symbol, the underscore can be omitted. The axes or the neighborhood area are set vertically downward and horizontally rightward, respectively.

**Shift Suffix**
There is a case in which operands of the target operation localize in the neighborhood area. For convenient notation, a PTB may have shift suffix indicating the amount of the origin offset. The shift suffix can be regarded as the address of the underscored position.

**Sum and Product of PTB’s**
In OAL, multiple results of product term operations are logically summed. This procedure is expressed by the sum of the PTB’s that correspond to the product term operations. The product of PTB’s is identical to a PTB whose elements are composed of the result of the logic product of individual functions at the corresponding position in the PTB’s. The product of PTB’s can be used for separating operands in a PTB.

For example, the following kernel expression specifies a logic expression, which is indicated by Eq. (A2).

\[
[10 \ 0.] + \begin{bmatrix} 1 \\ 0 \end{bmatrix}_{0,1}, \quad (A1)
\]

\[
c_{i,j} = a_{i,j-1}\overline{b}_{i,j-1}a_{i,j} + a_{i,j+1}\overline{b}_{i+1,j+1}. \quad (A2)
\]

Appendix B
A kernel expression of template matching for two images (Section 2) is

\[
[11 \ 00]. \quad (B1)
\]

Kernel expressions of inference with token propagation (Subsection 4.C) are as follows:
1. Horizontal marker propagation (see Fig. 9).
   Step (1) Vertical pattern expansion (input A, query (1); input B, don’t care):
   \[
   \prod_{i=-S+1}^{S-1} [1,1]_i; \quad [0.]. \quad (B2)
   \]

   Step (2) Creation of the condition image from the knowledge base (input A, knowledge base; input B, attribute plane of knowledge base):
   \[
   [11] + [,1 \ 10] + [,1 \ 10] + [.,1 \ 10]. \quad (B3)
   \]

   Step (3) Template matching (input A, expanded image (1); input B, condition image (1)):
   \[
   \sum_{i=0}^{3} \left( \begin{bmatrix} 00 \\ EE \end{bmatrix}_0 \cdot + \begin{bmatrix} 00 \\ EE \end{bmatrix}_0 \right). \quad (B4)
   \]

2. Change in the link direction pixels on the tokens (see Fig. 10; input A, query (1'); input B, don’t care):
   \[
   [11] + [,1 \ 10]. \quad (B5)
   \]

3. Horizontal token propagation (see Fig. 11).
   Step (1) Horizontal pattern expansion (input A, query (2); input B, don’t care):
   \[
   \prod_{i=-N+1}^{N-1} [1,1]_{0,2i}; \quad [0.]. \quad (B6)
   \]

   Step (2) Creation of the condition image from the knowledge base (input A, knowledge base; input B, attribute plane of knowledge base):
   \[
   [11] + [,1 \ 10]. \quad (B7)
   \]
Step (3) Template matching [input A, expanded image (2); input B, condition image (2)]:

\[
\sum_{i=0}^{3} ([11]_{0,-i} + [11]_{0,0-i}). \tag{B8}
\]

4. Change in the link direction pixels on the tokens [input A, query (2'); input B, attribute plane of knowledge base]:

\[
[01 \ 10]. \tag{B9}
\]

In the above expressions and equations, \(\Sigma\) and \(\Pi\) denote sum and product operations in kernel expressions, respectively; \(S\) is the number of subnetworks, and \(N\) is the number of nodes. The input images \(A\) and \(B\) in expressions (B2)–(B9) are represented by the image names in Figs. 9–12. The attribute plane of the knowledge base is shown in Fig. 6(f). In expressions (B2) and (B6), the symbol \(;\) means that the following operation is executed after the previous operation has been done.

References