Université de Neuchâtel
Institut de Microtechnique

Compact All-Optical Recurrent Neural Network

(Réalisation optique d'un réseau de neurones avec contre-réaction)

Thèse

Prèsentèe à la Faculté des sciences
pour obtenir le grade de docteur ès sciences

par

Christoph Berger

Neuchâtel, octobre 1998

IMPRIMATUR POUR LA THÈSE

Réalisation optique d'un réseau de neurones avec contre-réaction

de M. Christoph Berger

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel sur le rapport des membres du jury,

MM. R. Dändliker (directeur de thèse), H. Hügli, M. T. Gale (CSEM, Zürich) et J. Jahns (Hagen)

autorise l'impression de la présente thèse.

Neuchâtel, le 14 octobre 1998

Le doyen:

F. Stoeckli

Summary

We report on the design, construction and testing of an optical neural network with 256 fully interconnected neurons, reconfigurable weight matrix and optical feedback loop.

The building blocks of the optical system are discussed and the considerations that lead to the final design are presented. The principle of the interlaced fan-out is introduced. This method of interconnecting each channel with all other channels respects the complexity of different system stages in a better way than conventional methods. Imaging errors are minimized where the complexity is largest, i.e. at the interconnection matrix (weight matrix). In combination with a microlens array based image relay system (microchannel-telescope), large image fields with low aberrations over the whole field are possible. The potential of this approach is demonstrated with a matrix-vector-multiplier, where 65'536 individual beams all match their corresponding pixel apertures of a spatial light modulator that is used to introduce the interconnection matrix. The throughput of the system is sufficient for the implementation of an optical feedback loop, which is experimentally demonstrated. A constant channel pitch throughout the whole system results in a robust setup, which is easy to align.

A liquid crystal light valve acts as an array of neurons in our optical implementation of a neural network. The structure of the device and the physical properties relevant for its use as an array of neurons are presented. Experimental characteristics are given, and the problems which limit the performance of the device are described.

The performance of the complete optical neural network is limited by the weakest element in the chain, which is the liquid crystal light valve that suffers from a severe non-uniformity. The study of the network performance remained therefore of qualitative nature. Suggestions for improvements and for the future use of the system are given.

Table of contents

1 Introduction

1.1 Prologue and thesis overview

In spite of the enormous progress in computer technology during the last few decades, there are still numerous problems that are not satisfactorily solved by today's computers: classic examples are the real-time recognition of faces and voices.

Such tasks, however, are very well mastered by the human brain, which has a completely different structure than a computer: A computer serially processes information at a very high speed. The signal propagation in the brain is about 10^6 times slower, but the information is processed by a huge number (10^{10}) of simple processing elements, the neurons. These neurons are massively interconnected (~10^4 interconnections per neuron) and process incoming information in parallel.

The theory of neural networks evolved on one hand from the efforts to understand the data processing mechanisms of the human brain and on the other hand from the wish to technically simulate the processing power of our brain. Since the beginning of this theory in 1943 (cf. Appendix A), many different mathematical models have been developed. Of course, none of these models reaches the complexity and power of the information processing mechanisms of our brain. But for specific tasks, such artificial neural networks deliver useful results and are successfully used in commercial applications.

The technical implementation of massively parallel processing hardware is not a trivial task. The common method to implement artificial neural networks has been (and still is) to serially perform the large number of calculations related to neural networks on fast conventional computers. The enormous progress in computer technology allows for acceptable processing times, if the number of neurons is not too large.

Amongst the many possible ways to implement artificial neural networks, there is also the optical approach, which tries to benefit from the inherent parallel processing power of optics: In contrast to classical electronics, optical signals can propagate through the free space. In combination with suitable 2D emitter-, modulator- and receiver-arrays, this allows for the construction of three-dimensional processing systems. Together with the fact that there is no direct interaction (crosstalk) between optical beams, this leads to a massive increase in achievable parallelism and data bandwidth. Further advantages of optics over electronics include immunity towards electromagnetic influences and much easier impedance matching.

The advantages of optics in information processing are of course not only useful for optical implementations of neural networks. Currently, a considerable effort is being made worldwide to replace bandwidth-limited electronic interconnections between processing elements (e.g. between silicon microchips or between building blocks of a computer) by optical interconnects.

Be it for neural networks or be it for interconnects, optics is in both cases an attractive approach. Of course, the fields are closely related, because a central characteristic of neural networks are the massively parallel interconnections between their processing elements. Even if purely optical neural networks might not be the solution of the future (cf. section 1.4), neural network applications will still benefit indirectly from optics thanks to the expected increase of computational power of data processing systems using parallel optical interconnects.

This thesis is a an overview of the author's contributions to the field of optical information processing. They include a new concept for fully interconnecting 2–dimensional arrays of processing elements (interlaced fan-out), and the successful experimental demonstration of the microchannel-telescope image relay principle.

The design, construction and testing of a recurrent optical neural network has been performed during the last four years at the Institute of Microtechnology of the University of Neuchâtel, Switzerland. Part of the work has already been published and presented at international conferences:

[A] N. Collings and C. Berger, *Demonstration and discussion of an interlaced fan-out interconnect*, Inst. Phys. Conf. Ser. **139**: Part II, 247-250 (1994).

[B] C. Berger, N. Collings, A. Pourzand and R. Völkel, *Comparison of Two Reconfigurable NxN Interconnects for a Recurrent Neural Network*, Opt. Rev. **3**, No. 6A, 388-390 (1996).

[C] C. Berger, N. Collings and T. Jost, *Recurrent Optical Neural Network for the Study of Pattern Dynamics*, in "Optics in Computing", 1997 OSA Technical Digest Series **8**, 46-48 (1997).

[D] C. Berger, N. Collings, R. Völkel, M. Gale and T. Hessler, *A microlens-array-based optical neural network application*, Pure Appl. Opt. **6**, 683-689 (1997).

[E] C. Berger, N. Collings and D. Gehriger, *Recurrent Optical Neural Network for the Study of Pattern Dynamics*, in "Optical Memory & Neural Networks", Proc. SPIE **3402**, 233-244 (1998).

[F] C. Berger and N. Collings, *All-Optical Recurrent Neural Network*, in "Optics in Computing", Proc. SPIE **3490**, 465-469 (1998).

The first chapter of this thesis (Introduction) gives a short introduction to artificial neural networks and their optical implementation. An overview of the field and the context of this work are presented.

The presentation of our neural network is split up in three parts: The neurons, the interconnections and the complete network. In our optical approach, the non-linear signal processing of the neurons is performed by a liquid crystal light valve (LCLV).

The whole rest of the optical system represents a reconfigurable full interconnect between the output and the input of all the neurons. In addition, the optical system is responsible for initialization and monitoring of the network.

In chapter 2, the discussion is started with the optical system, because this has been the main part of the work. Various approaches for the several building blocks of the system are discussed, and the considerations that led to the design of the final setup are presented. We introduce a new concept, the interlaced fan-out and demonstrate the potential of the method in a matrix-vector-multiplier using microchannel-telescopes. With this approach, that minimizes imaging errors for systems with a large number of channels, it has been possible to realize an optical neural network with 256 neurons, 65'536 individually reconfigurable connections links and optical feedback. The optical test of the setup shows that the imaging quality satisfies the demanding requirements of a fully interconnected system and that the setup offers enough throughput for optical feedback.

The liquid crystal light valve (LCLV), which acts as an array of neurons, is presented in chapter 3. The structure of the device and the physical properties that are relevant for its use as an array of neurons are presented. Experimental characteristics are given, and problems that limit the performance of the device are described.

In the fourth chapter, results of experimental work with the complete neural network are presented.

Chapter five closes this thesis with a summary of the work and some conclusions. An outlook to possible future studies with the present setup is provided.

Additional background information is provided in the appendices. Appendix A gives a short introduction to the fascinating field of neural networks and Appendix B summarizes some optics background relevant for this work. Microlens arrays have been very important components for this work. A summary of the technology and a comparison of different types of microlens arrays is given in Appendix C. Finally, detailed information on how to align the optical setup is provided in Appendix D.

1.2 Artificial neural networks

This section shall give a very rough overview of the fundamental elements of artificial neural networks. The goal of this section is to provide the necessary information for the understanding of our optical implementation of a neural network. A more detailed introduction to neural networks is provided in Appendix A.

Neural networks process information with a large number of simple processing units, which are called *neurons* (also units, cells or nodes). The neurons can pass signals between each other over directed *connection links*. Every neuron has one sender element, the *axon*, over which it can send a binary signal to other neurons. To receive signals from other neurons, every neuron has several receiving elements, the

dendrites. At the interface between the axon of one neuron and the dendrite of another neuron is the *synapse*, which can have inhibitory or excitatory properties. Mathematically, this is expressed by attributing a *weight* to every connection link, which is typically used to multiply the transmitted signal. The neuron collects and sums up the weighted signals from all its dendrites. If the summed signal is strong enough (above a certain *threshold*), the neuron "fires", i.e. it sends a signal via its axon to other neurons. Mathematically, this process is described by applying a nonlinear function (called *activation function* or *threshold function*) to the sum of the weighted neuron inputs.

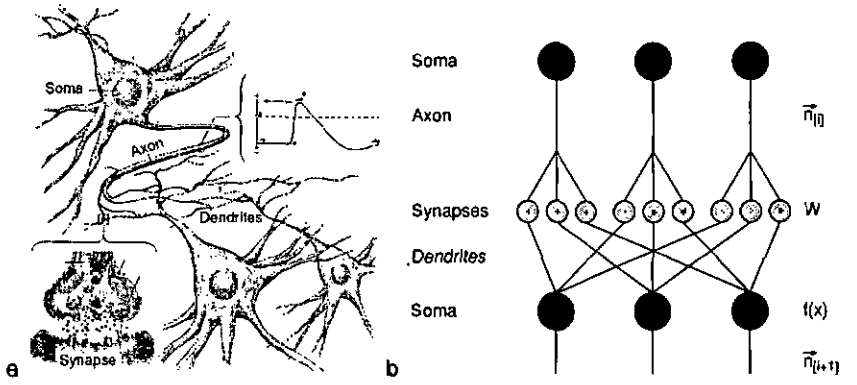


Fig. 1.1: a) Signals are exchanged between neurons by means of electrical pulses, generated by the cell body (soma). At the interface between the sender element (axon) of one neuron and the receiver element (dendrite) of another neuron is the synapse (blow-up lower left), where the signal is excited or inhibited by electrochemical processes (From [1]). b) Schematic sketch of an artificial neural network (cf. Eq. 1-1).

A given vector of input signals will propagate through the network of neurons and synapses and produce a vector of output signals. The kind of operation that is performed in this process depends on 1) the pattern of connections between the neurons (called the *architecture* of the neural net), 2) the ensemble of information stored in the synapses, and 3) the threshold function of the neurons (which is normally identical for all the neurons of a net).

The choice of a specific neural network model for a particular task normally determines the architecture and the threshold function. The task-specific information or system-knowledge is obtained by means of learning algorithms and is stored in the ensemble of all the synapses (or weights). Mathematically this information can be expressed in the form of a suitable matrix, which is called *weight-matrix*. The passage of signals from a number of neurons via the synapses to other neurons can therefore be expressed by the multiplication of an input vector with the weight-matrix. The number of elements of the input vector corresponds to the number of input

neurons and the number of elements of the weight-matrix corresponds to the number of synapses.

In a multi-layer network, signals are fed forward to other neurons. In other models, like the Hopfield-model that has been used for this work, signals are fed back to the same set of neurons. To describe the evolution of the state of a network with feedback, an iteration-number i is introduced. The evolution of the network state can then be described as

$$\vec{n}_{[i+1]} = f(W \cdot \vec{n}_{[i]}) \quad , \tag{1-1}$$

where $\vec{n}_{[i]}$ is a vector describing the state of all the neurons in the i-th iteration, W is the weight matrix and f is the threshold function (cf. Fig. 1.1).

The non-linear threshold function can be a binary function like

$$f(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \le 0 \end{cases} \quad , \tag{1-2}$$

or a sigmoid-function like

$$f(x) = \frac{1}{1+e^{-cx}} \quad . \tag{1-3}$$

The elements of the weight matrix W can be calculated in a way that they represent a certain information, e.g. a set of reference patterns (or vectors). If a test pattern similar to a reference pattern is presented to the network, the net will after some iterations converge to a stable state that represents the corresponding reference pattern.

More details about the Hopfield-model can be found in Appendix A.3. For the following discussion it is just important to retain that the two central operations in the Hopfield-model are a matrix-vector-multiplication and the application of a non-linear threshold function.

1.3 Optical Implementation of neural networks

The optical implementation of neural networks was motivated by the massively parallel nature of neural networks and the potential advantages of optics for such tasks. Apart from the details already provided in section 1.1, optics represents an obvious approach for image related tasks like pattern recognition, etc. due to the 2D nature of this data.

The paper by Psaltis and Farhat [2] is generally considered to be the first publication introducing the idea of the optical implementation of neural networks. Their work based on Hopfield's model [3] and on a publication by Goodman *et al.* [4] who described an optical implementation of a matrix-vector-multiplier.

The input vector in this classic version of the optical matrix-vector-multiplier (cf. Fig. 1.2) is represented by a 1D line of N light emitting diodes (LEDs). The information coded in this vector is distributed (or fanned-out) to the elements of the weight matrix (synapses) with a telescope consisting of a normal lens and a cylindrical lens. The light of each input LED is thus smeared to one column of the weight matrix. The weight matrix consists of a mask of $N \times N$ cells with different transparency. After passage of the light through the weight matrix mask, the information of each matrix row is collected (fanned-in) by a similar type of optics (rotated by 90° with respect to the input optics) and focused on a 1D line of N detectors.

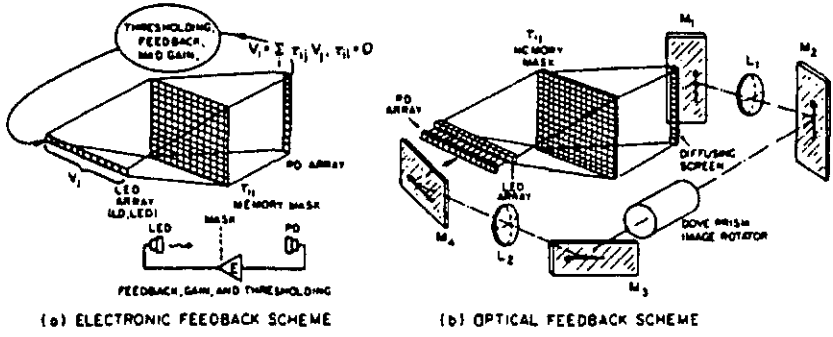


Fig. 1.2: Matrix-vector-multiplier with electronic (a) and optical (b) feedback. Scheme from the first paper proposing the optical implementation of a neural network [2].

To allow for a higher number of channels and for applications dealing with image data, the elements of the input vector can be arranged in a 2D array. This of course changes the arrangement of the elements in the weight matrix and therefore also the optics required to route the light signals from the input via the weight matrix to the output (which is now a 2D array of detectors). The present work falls in this category of arrangements.

As for the feedback, Psaltis and Farhat already proposed two schemes: one with electronic feedback and one with optical feedback (cf. Fig. 1.2). Their optical feedback, however, still required a block of electronics (detectors, connecting electronics and emitters) in order to apply the nonlinear threshold function and to provide the gain that is needed to compensate for the losses of the feedback-loop.

With the availability of optically addressable spatial light modulators (SLMs) with gain and non-linear properties (like e.g. liquid crystal light valves, LCLVs), the possibility for real all-optical feedback was given. However, the gain of these devices has been

rather limited and therefore most of the experimental implementations were restricted to only few channels or they used electronic feedback.

Apart from the high signal propagation speed and the continuous update of information, optical feedback is interesting because of the potentially asynchronous processing properties of optically addressed SLMs (cf. section 3.7). Asynchronous update of the neuron states provides better performance than synchronized operation [5, 6]. And certainly asynchronous operation corresponds much more to the information processing mechanisms of our brain. Of course, asynchrony can also be realized electronically, but this requires an independent threshold-and-gain-module per neuron, whereas in optics a single 2D element (as a LCLV, for example) can do the job.

A large number of different optical implementations for all kind of neural networks has been proposed and partially realized since 1985. A collection of selected papers on optical neural networks can be found in [7]. [8] is a more recent overview, but unfortunately the author concentrates only on implementations using photorefractive optics.

1.4 Current international activities

While this project was carried out, a decrease of the international activities in the domain of optical neural networks (indicated by a decreasing number of conference contributions and publications) could be observed.

We believe that this trend is, amongst other factors, due to a lack of powerful demonstrators where optical neural networks could really demonstrate their benefits in real-world applications. A lot of concepts have been proposed, but the experimental demonstration has often been limited to a small number of neurons. Very often, researchers who want to build an optical neural network are faced with optical and electro-optical hardware that is still in a prototype phase and that is usually not sharing common standards (e.g. channel pitch), which results in non-ideal setups. Experimental demonstrators are mostly voluminous, delicate to align and expensive.

On the other hand, one can already buy moderately priced handheld devices that base on standard (serial) computing algorithms or on special neural network silicon chips and that successfully implement neural network functionality. Fig. 1.3 shows an example of a 15 cm long device that performs scanning, character recognition and translation of the recognized word. The reason for the lead of software and silicon hardware solutions over optical solutions is twofold: They offer more flexibility in adapting an algorithm to a specific task and they use relatively cheap standard mass-produced hardware.

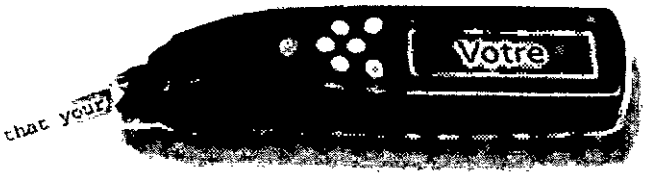


Fig. 1.3: "Quicktionary™", a handheld device that performs scanning, character recognition and translation. A specialized neural network chip supports the character recognition.

Respecting the fact, that research on domains with promising applications has better chances for the necessary funding, many groups have turned their attention to other topics, which are related to the still growing computer industry. Key domains of interest include optical interconnects and holographic storage. Many of these people believe that optics is a promising *carrier* of information but that the information *processing* should be done by the established digital semiconductor technology. Remarkably, a main international annual meeting in the field has changed its name from "Optical Computing" to "Optics in Computing".

The largest active domain where optics is still used for *processing* information is probably the research on correlator systems.

Future progress in micro-optical system technology (e.g. motivated by the push towards optical interconnects in computer systems) may result in cheap, standardized and easy combinable micro-optical and opto-electronic elements. The availability of such hardware may eventually re-motivate people to construct optical neural networks. A key factor for the future of optical neural networks will surely also be the existence of attractive and demanding applications (large number of neurons) that cannot be satisfactory solved by other approaches.

1.5 History and context of this work

Optical neural networks have been studied at the Institute of Microtechnology since 1990. Several prototype optical neural networks have been constructed and characterized, including hybrid and all-optical Hopfield networks, and multi-layer networks [9-11]. Another main activity in the domain of optical computing concentrated on correlator systems [11]. Besides the system activities, IMT has numerous ongoing projects on micro-optical device technology, such as diffractive elements, retractive microlens arrays, micro-opto-mechanical elements, liquid crystal devices, etc.

This project evolved from the optical systems discussed in the Ph.D. works of two former collaborators, Weible and Xue. Weible [9] realized three different types of optical learning systems, each employing liquid crystal devices and diffraction gratings. The first type was an associative memory, realized as a fully interconnected single-layer Hopfield neural network. An all-optical solution with 3 x 3 channels was used to implement an inhibitory Hopfield network, and a hybrid solution (thresholding and feedback by means of a computer) with 7 x 7 channels was used to implement an inhibitory and an inverted Hopfield model. The use of a fixed, pre-calculated weight matrix was compared to an in-situ trained weight matrix using a Hebbian type of learning algorithm. The in-situ trained version that could compensate for a part of the system imperfections produced clearly better results. The two other types of learning systems were a self-organizing Kohonen neural network tested with the "travelling salesman problem", and an optical parallel processor used for the optical optimization of binary phase gratings.

In the first part of his work, Xue [10] focuses on the characterization of different types of liquid crystal light valves. In the second part, he reports on the all-optical implementation of an inhibitory and an inverted Hopfield neural network with 3 x 3 channels. Two categories of system behavior were observed, independent of the model: Fast convergence of the system output to one stable state and a slow evolution to a regular oscillation between two states. Fig. 1.4 shows an example of such an oscillation (see the figure caption for more details).

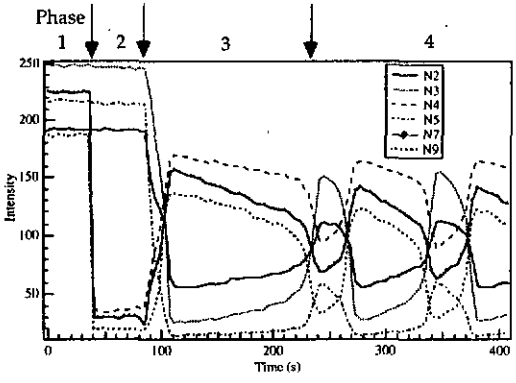


Fig. 1.4: Oscillation between stable states, observed in a 3 x 3 recurrent optical neural network [10]. Each line represents the state of one neuron. Phase 1: No initial input, feedback loop closed, all neurons high (non-uniformity is due to system limitations); phase 2: with initial input, feedback loop closed; phase 3: feedback loop opened, transition phase; phase 4: oscillation between two stable states (cycle time about 100 s).

Oscillatory attractors have also been observed in electronic analog neural networks by Marcus *et al.* [12] and in optical neural networks by Lee *et al.* [13]. Xue

compared these results with his system and concluded that the observed oscillations originate from the ratio between response and propagation time (10^7), from the negative connections of the inhibitory and the inverted Hopfield model and from the dilute weight matrix. From these considerations, the system should theoretically continue to oscillate. The observed existence of a stable state was found to be a consequence of the non-uniform response time of the neurons in the optical system. The study of these phenomena was limited by several factors such as the small number of neurons, a fixed weight matrix (which cannot adapt for system imperfections), low efficiency and insufficient uniformity of the available optical hardware.

Progress in device technology motivated us to demonstrate a more complex and at the same time more compact Hopfield neural network with all-optical feedback. The dynamics of patterns circulating in such a network, as observed by Xue, should be studied after completion of the optical setup.

In the context of what has been said in section 1.4, the neural network aspect has become somewhat secondary during this work. The main focus shifted towards the experimental integration of novel micro-optical and electro-optical technology into a demanding and massively parallel interconnected system.

2 Design considerations and test of the optical setup

2.1 Overview and description of the task

2.1.1 System design goals

This work was the logical continuation of previous projects performed at our institute, as mentioned in section 1.5. The neural networks with optical feedback realized by Weible [9] and Xue [10] were *limited to 9 neurons* due to the optical components available. One of the goals of this work was thus to realize a recurrent all-optical neural network with a *significantly increased number of neurons* (or channels).

Another restriction of the previous systems was the need to work with a *fixed weight matrix* on a photographic mask. The gain of the available liquid crystal light valves was not high enough to permit the use of reconfigurable but highly absorbing liquid crystal televisions in the optical feedback loop. Another goal of this project was thus to build a system *with reconfigurable weight matrix.*

A third goal of this project was to build a rather *compact system* in order to get away from the cliche of optical data processing systems that fill up entire optical tables. The recent progress in microlens array fabrication motivated us to demonstrate a complex system using micro-optics.

2.1.2 Chapter overview

In the following sections, separate building blocks of the optical setup will be discussed. In order to make this report interesting for a wider audience, the discussion is kept as general as possible.

The determination of the final parameters of our system was a dynamically evolving process. The linear way of describing the system design considerations in this report will therefore not always reproduce the way the system evolved.

In section 2.2, some factors that limited the design freedom will be presented. The generation of a 2D input signal is discussed in section 2.3. The transport (or relay) of this signal to a next stage is discussed in section 2.4. The exchange of information between all the channels of an optical data processing system is the subject of

section 2.5. In section 2.6, all the building-blocks are put together to a complete system, whose optical properties are presented in section 2.7. Finally, some ideas for a future optimization of the optical setup are reported in section 2.8.

2.1.3 Preview on the final setup

The final setup consists of four main parts: initialization, matrix-vector-multiplier, neuron-plane and feedback (Fig. 2.1).

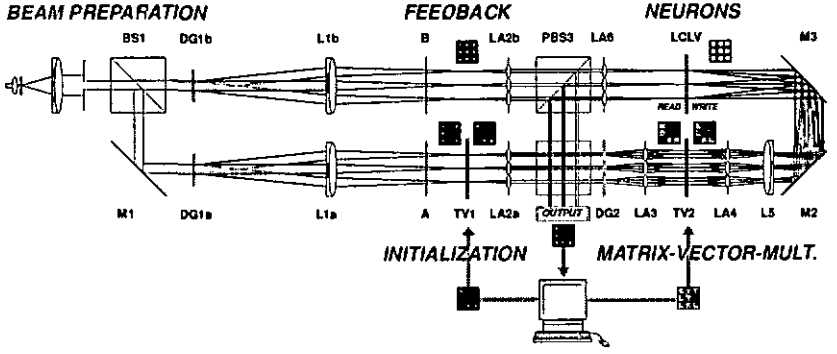


Fig. 2.1: Schematic preview on the final setup (cf. Text).

For initialization of the system, a 16 x 16 spot array is generated by means of a Dammann-grating (DG1a) and an achromat (L1a). The input vector (256 elements, arranged in an array of 16 x 16 channels) is introduced by means of a liquid crystal television screen (LCTV1), which is operated in amplitude modulation mode. A lens array is used to collimate the light coming from LCTV1. By means of another Dammann-grating (DG2) and a lens array (LA3), each individual input channel is replicated 16 x 16 times and imaged onto the second liquid crystal television screen (LCTV2), where the weight matrix information is introduced (256 x 256 matrix elements, arranged in 16 x 16 sub-arrays). By means of a lens array (LA4) and an achromat (L5), the modulated signal is fanned-in and imaged onto the write side of a liquid crystal light valve (LCLV), which acts as an array of neurons (neuron plane).

By means of the shutters A and B, the system can be switched from initialization to feedback mode. As above, 16 x 16 collimated beams are generated by DG1b, L1b and LA2b. After passage through a polarizing beamsplitter (PBS3), these beams are focused onto the reflective readout side of the LCLV by means of lens array LA6. The polarization state of the reflected beams is modulated with the (thresholded) information applied to the write side of the LCLV. The higher intensity on the readout side results in an effective signal amplification which is needed for the optical

feedback. The polarizing beamsplitter performs the transformation to amplitude modulation and feeds the signal back to the matrix-vector-multiplier. The information is now circulating in the feedback loop. Each channel is updated individually and asynchronously due to the asynchronous processing properties of the LCLV. In any mode, the state of the system can be monitored at the output of beamsplitter BS2.

2.2 Fixed parameters and boundary conditions

Boundary conditions and limiting factors like cost, size, time, availability of components, etc. strongly influence the design of optical systems. The presented setup was actually designed around some of the available electro-optic devices. These boundary conditions shall be presented in the following.

2.2.1 Lateral geometry and number of channels

A key element for the system design process were the liquid crystal television panels (LCTV), which are used to spatially modulate the light in our system.

Some years ago, high resolution LCTVs could only be found in videoprojector systems. The normal way to get such panels was thus to buy a videoprojector, dismount it and use the LCTVs with the original driver electronics. In the first years of commercially available videoprojectors, such systems were only used for continuous images like TV and VHS video signals. It was normally not possible to independently address individual pixels of these panels, and considerable efforts were made to overcome this restriction for optical data processing applications [11]. With the upcoming of mobile computing and multimedia, there was suddenly a demand for projectors that were able to correctly display discrete computer-generated images. One of the first commercially available models was the LitePro 580 from InFocus, which we bought in the early days of this work.

This projector has three LCTV panels (one for each basic color) with 640 x 480 pixels each. The pixels have a pitch of 42 µm x 42 µm and a pixel size of 21.5 µm x 31.5 µm. Fig. 2.2.a shows a photo of such a panel and Fig. 2.2.b shows an enlarged view of the pixels of this panel with only one individual pixel activated. The addressing of the three panels is made with the red, green and blue component of a VGA video signal.

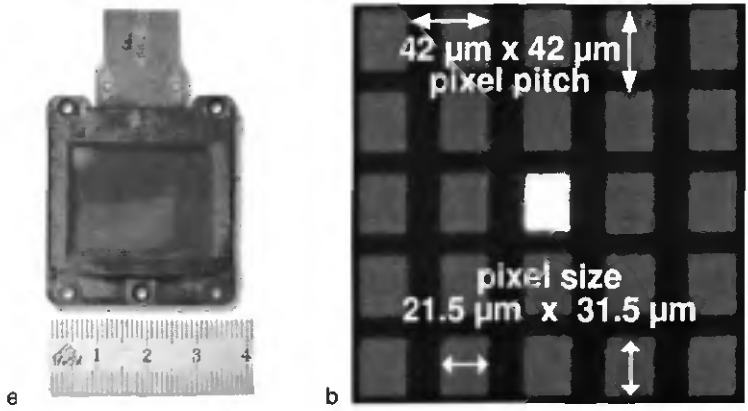


Fig. 2.2: a) Photo of a liquid crystal television panel (LCTV); b) Details of pixels and demonstration of an individually addressable pixel.

The pitch of the LCTV pixels determined the absolute size of the other system components. The number of pixels allows, in principle, for a fully interconnected system of 21 x 21 channels. However, to avoid crosstalk (cf. section 2.6.2), we wanted to have some space between the channels and we decided for *16 x 16 channels with a pitch of 26 pixels (= 1092 µm)* between the channels. This choice determined the lateral geometry of our microlens arrays (cf. Appendix C).

A word to the square number of channels: Due to the dimensions of the LCTV we could have built a system with a rectangular geometry and 50 % more channels (16 x 24). The restriction to a square number of channels and thus to a square geometry of our components proved to be very useful for the testing of the numerous components. With the square geometry, we were able to rotate individual components by 90°. This possibility often helped to isolate an element that was causing a problem.

2.2.2 Opto-mechanics and minimal focal length

Another important decision was to base our setup on commercially available mounts. After initial tests, we decided for the German OWIS-65 system, which is a rail-based system with a relatively low optical axes (65 mm above breadboard level). The basic building blocks of this system are normally 65 mm x 65 mm wide and 20 mm thick. This thickness defined a lower limit for the distance between the optical elements of our setup and thus for the focal length of some elements.

2.2.3 Wavelength

The wavelength was a fixed parameter. For availability reasons, the 488 nm line of an Ar^+-laser has been chosen as the operating wavelength.

2.3 Generation of input vector spot array

The input vector of our neural network has 256 elements, which are arranged in an array of 16 x 16 spots. The task is thus to generate an array of 16 x 16 beams and to modulate the input information onto these 256 beams. This kind of task is found in a lot of parallel information processing optical systems.

Three different concepts have been tested: 1) illumination of a microlens array, 2) combination of a diffraction grating and a lens and 3) an array of vertical cavity surface emitting lasers (VCSEL).

For the first two methods, a spatial light modulator (SLM, e.g. a LCTV) is needed to introduce the input information. Size and position of the SLM pixels define the required geometry of the spot array incident on the SLM. After comparison of different fan-out methods (cf. section 2.5), we decided for an interlaced fan-out (cf. section 2.5.2) with a widely spaced input array (dilute array). The pitch of the input channels is in our case 1092 µm (cf. section 2.2.1), the side length of the whole input image is thus 15 x 1092 µm = 16.38 mm. Over this whole area, the lateral position of the generated spots had to be precise enough to match the position of the corresponding LCTV pixels. Considering the size of one pixel aperture (21.5 µm x 31.5 µm), we required a precision of ±10 µm for the lateral position of the generated spots (over the whole image field). The dilute input array may be a specialty of our setup, but when the number of channels really gets massively parallel (e.g. 1024 x 1024), even a small channel pitch will produce a large image field. Viewed in this context, the considerations concerning the required imaging precision are of a more general interest.

When comparing the different methods, not only the quality (i.e. spot size and array geometry) of the spot array *in the input plane* has to be considered, but also the behavior of the beams *afterwards* (e.g. overall divergence of the set of beams) with respect to their further use in the system.

2.3.1 Illumination of a microlens array

An array of N x N microlenses is illuminated by an expanded and collimated laser beam (Fig. 2.3, N = 16 in our case). The main advantage of this approach is that the lateral position of the generated spots is given by the quality of the microlens array (details about the tested microlens arrays are given in Appendix C). Another advantage is the simplicity of this arrangement. The quick and easy alignment made it a favorite method for the testing phase. The drawbacks of this simple method are non-uniform illumination due to the Gaussian intensity profile of the incident beam and diffraction related problems (see below).

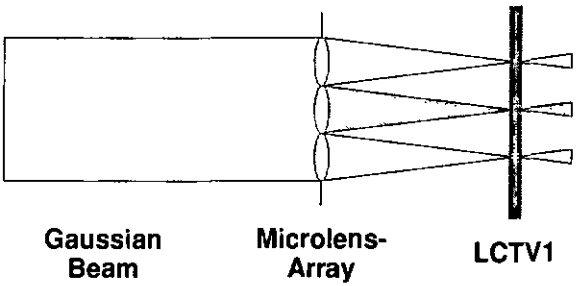


Fig. 2.3: A quick method to generate a spot array is to illuminate a microlens array with an expanded beam. The main advantage is the well defined geometry of the generated spot array; disadvantages include non-uniform illumination and diffraction related problems like bad throughput and crosstalk.

There are several concepts to improve the uniformity of the illumination: In the simplest configuration, only the central part of a widely expanded Gaussian beam is used (Fig. 2.4.a). Another method is the use of a grayscale filter which compensates for the Gaussian intensity distribution within the beam (Fig. 2.4.b). Both methods have the disadvantage that a lot of the available laser power is wasted, which is acceptable for the testing of elements and building blocks, but can be a major problem in a large system (e.g. our system in feedback mode). We also thought about using a beam shaping element for better illumination uniformity [14]. Such diffractive elements can transform a Gaussian beam profile into e.g. a square profile with uniform (< ± 5 %) intensity. The efficiency of such a device is typically about 80 % to 85 % (8 level-design). However, the new profile is only generated in the Fourier plane of a lens. The element is therefore not suited for use in a multistage system.

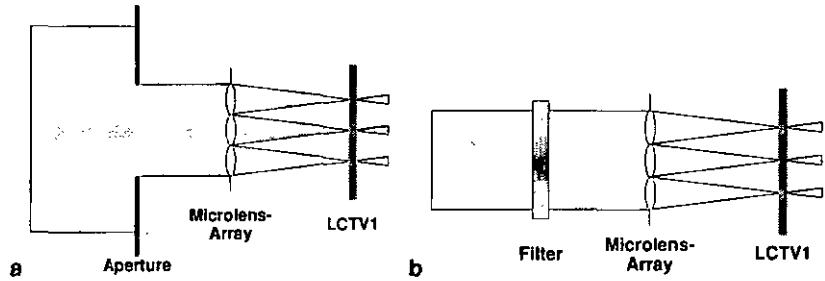


Fig. 2.4: Two concepts to improve the uniformity of the illumination.

The problems of bad throughput and crosstalk are related to the laws of diffraction that impose a lower limit to the achievable diameter of the focal spot of a lens [15, chapter 10, 16, section 4.3] (cf. Appendix B.2). For a lens with focal length f and a square aperture of side length D, the focal spot has a square shape of minimal side length d, which is given by

$$d = 2\frac{\lambda f}{D} \quad . \tag{2-1}$$

In our case ($\lambda \approx 500$ nm and $D \approx 1$ mm), we get $d \approx 10^{-3} \cdot f$. For our specific situation, we can therefore derive a rule of thumb: The diffraction limited spot size of a particular microlens in micrometers is about equal to the focal length of this lens in millimeters. Due to the thickness of the optical mounts, we worked with relatively long focal lengths between 5 mm and 80 mm. The corresponding diffraction limited spot diameters were thus between 5 µm and 80 µm. If we compare this to the size of the pixels of the input LCTV, we see immediately that there is a conflict for $f > 20$ mm. If the spot size is larger than the pixel aperture, we a) introduce losses and b) cause diffraction at the pixel aperture, which results in crosstalk between channels. Fig. 2.5 illustrates how light from one specific input channel is not only captured by the corresponding microlens of LA2 but also by the neighboring microlenses. Geometrical considerations (Fig. 2.6) show that light coming from one specific channel moves into the next channel after a distance that corresponds to *f2*. For distances smaller than N (= number of channels) times *f2*, we observe a superposition of "ghost-patterns" to the original pattern. For distances larger than N times *f2*, the correct pattern is captured by LA3.

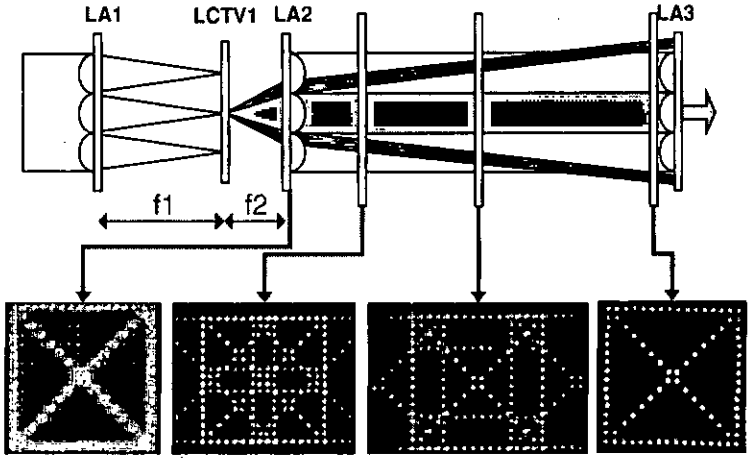


Fig. 2.5: Effect of diffraction at the LCTV pixels I: Crosstalk between channels. Due to the spread of the light cone, light from one channel can be captured by the lens of another channel. For distances smaller than N times f2, this results in a superposition of "ghost-patterns" to the original pattern (cf. Fig. 2.6 and text).

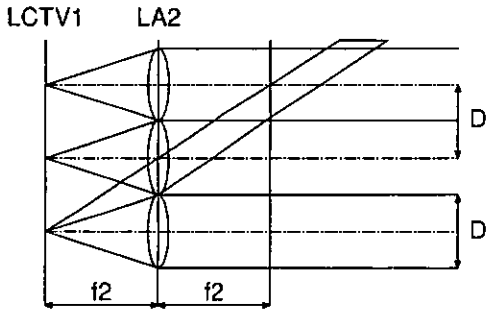


Fig. 2.6: Effect of diffraction at the LCTV pixels II: Light coming from one channel moves into the next channel after a distance f2.

At a first glance, the solution of this problem seems to be easy: just use shorter focal lengths (thinner holders) to capture the complete light cone emerging from LCTV1. But this is only true for an isolated setup or at the end of an optical system (e.g. in the detector plane). At the input of a multi-stage system, however, a shorter focal length of the illuminating microlens array can have very undesired consequences.

To understand this, we look at the beams after passage through the LCTV. The diverging beams have to be captured by a telescope, which relays them to a further stage of the system. With a constant microlens aperture (given by the channel pitch),

a decrease of the focal length results in an increase of the divergence. In order to catch all the light emerging from LCTV1 and to avoid crosstalk between the channels, we have to decrease the focal length of the input lens of the following telescope (Fig. 2.7). For different telescopes (cf. section 2.4), this has different consequences: For a 4f–telescope (Fig. 2.7, top), a shorter focal length of the telescope lenses (at a constant lens diameter) results in larger imaging errors. For a microchannel-telescope (Fig. 2.7, bottom), the length L of the region without crosstalk decreases. Moreover, the available space for components within the telescope (e.g. feedback optics) is reduced in both cases. A detailed discussion of the telescopes follows in section 2.4.

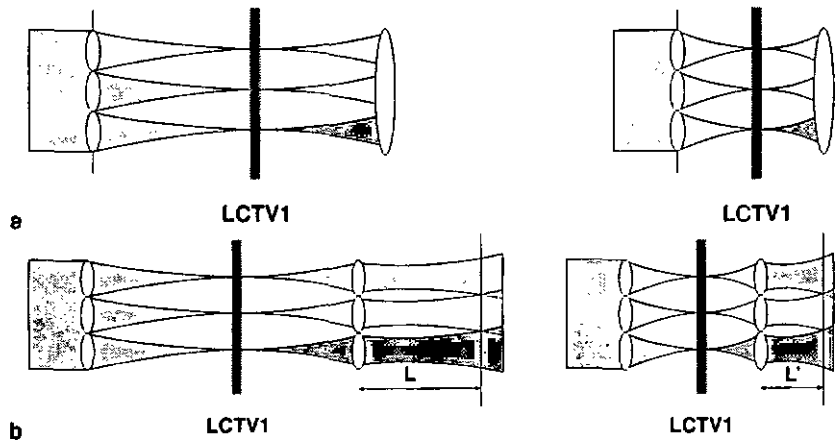


Fig. 2.7: Decreasing the focal length of the illuminated microlens array helps to reduce the size of the spots incident to the LCTV, but the focal length of the input lens of a following telescope has to be reduced as well in order to catch all the light and to avoid crosstalk. In consequence, the available space, e.g. for a beamsplitter, is reduced and the influence of aberrations is increased.

2.3.2 Fan-out with grating and lens

A second method of generating a spot array on the input LCTV is to use a diffraction grating [17, 18] and a lens (Fig. 2.8).

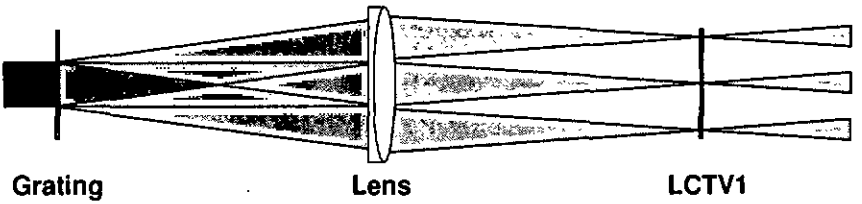


Fig. 2.8: A diffraction grating in combination with a lens can be used to generate an array of spots. The uniformity of the intensity distribution depends on the quality of the grating. The diameter of the illuminating beam determines the spot size.

In contrast to the illuminated microlens array, the uniformity of the intensity distribution depends only on the kind and quality of the grating. No light is wasted and the size of the generated spots can be controlled by changing the diameter of the incident beam. Drawbacks of this arrangement are larger size and potential problems with imaging errors (see below).

In the following, several aspects of this arrangement will be discussed, starting with the position of the grating with respect to the lens.

e) Grating position I: grating before lens

If a grating with periodicity Λ is placed in front of a lens with focal length f, the pitch Δ of the generated spots is given by:

$$\Delta = n \cdot f \frac{\lambda}{\Lambda} \quad , \tag{2-2}$$

where λ is the wavelength and n is the diffraction order (cf. Appendix B.5). In our case, $n = 2$ because we used a grating with "even-orders-missing" design [19, 20]. As we can see from this formula, the pitch of the spots in the focal plane does *not* depend on the exact position of the grating. This becomes immediately plausible if we look at the lens as a Fourier-transformer that transforms angular information into spatial information [15, p. 504]. However, the *angle of incidence* of the beams that form the spot array is strongly influenced by the axial position of the grating (Fig. 2.9).

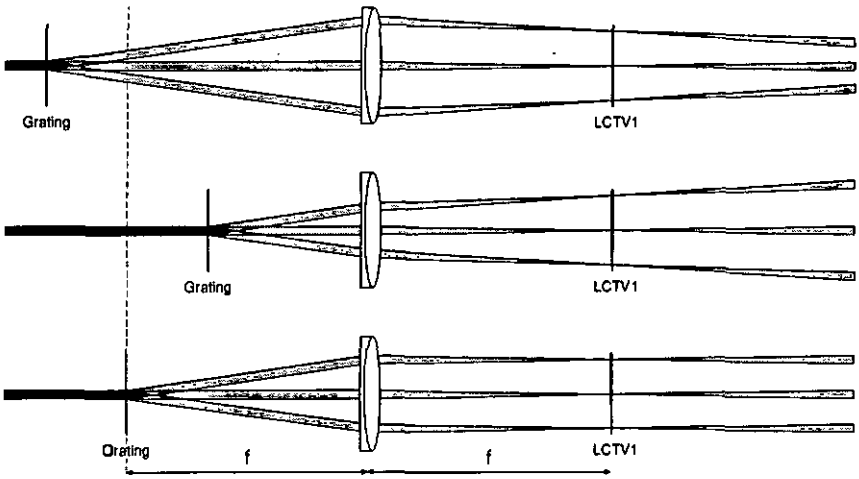


Fig. 2.9: A grating is placed in front *of a lens in order to generate a spot-array on the input LCTV. The pitch of the spots in the focal plane is independent of the axial position of the grating, but the angle of incidence changes. For a system with several stages, the best axial position for the grating is the front focal plane.*

If one just has to produce a spot array, the grating can be placed very close to the lens in order to get a compact system. In a system with several stages, however, the grating should be placed in the front focal plane. This results in perpendicular incidence of the principal rays and thus in the smallest overall divergence of the set of beams after the LCTV.

A potential problem of this configuration is the fact that reasonably priced commercial glass lenses have a *focal length tolerance* of about 1 % (according to the datasheets of several manufacturers). Because the pitch Δ is directly proportional to f, we have to expect deviations of the same order of magnitude. In our case, the pitch is 1092 µm, the potential deviation therefore 10.92 µm. For a point at the edge of the array, this can result in a deviation of up to 82 µm from its desired position on the LCTV (if the image is correctly positioned in the center). This is 8 times more than the 10 µm tolerance that are required to have the spot within the LCTV pixel aperture.

One way to overcome this problem is to use a pair of lenses instead of a single lens, which allows for an adjustable effective focal length. The alignment of such a system, however, is rather tricky because the position of front and back focal plane changes when the effective focal length is adjusted. [11, 21, section 2.10]

b) Grating position II: grating after lens

On the other hand, the grating can also be placed between the lens and the LCTV. In this case, Eq. (2-1) becomes

$$\Delta = 2 \cdot z \frac{\lambda}{\Lambda} \quad , \tag{2-3}$$

where z is the distance between grating and LCTV. The big advantage of this method is that the pitch can be adjusted by moving the grating along the z-axis. The divergence of the whole set of beams, however, is not ideal (Fig. 2.10). Moreover, the generated spots do not lie within the LCTV plane but on a spherical surface with radius z, centered at the grating. Compared to above, larger diffraction angles and thus smaller feature sizes of the grating are required.

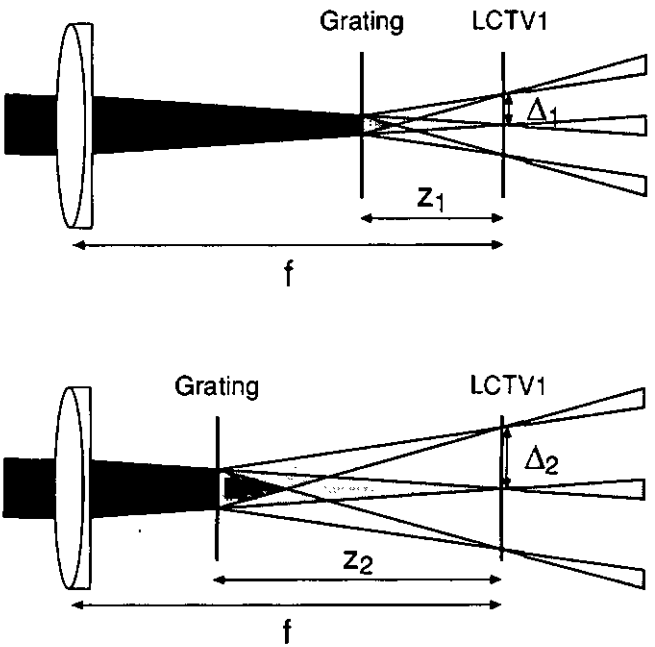


Fig. 2.10: A grating is placed behind a lens in order to generate a spot-array on the input LCTV. The pitch Δ of the spots can be adjusted by moving the grating along the z-axis. The angle of incidence is not perpendicular and the generated spots do not lie in the LCTV plane but on a spherical surface with radius z, centered at the grating.

c) Imaging errors

A general problem in matching a large array of spots to a large array of pixels are imaging errors caused by aberrations. A short summary of this topic and references to further information are given in Appendix B.3.

Due to the large field of view of our setup and the rather strict requirements for the alignment precision, imaging errors have been an important factor throughout this work. We identified field curvature as a main issue. A more detailed discussion will follow in section 2.4 (telescopes).

As for the spot array generation, we looked at several lenses using a raytrace software [22]. The result of this study was that in our case only lenses with a focal length ≥ 300 mm produce acceptably small imaging errors (the amount of Petzval curvature is inversely proportional to the focal length, cf. Appendix B.3). With the Dammann-grating positioned in the front focal plane of the lens, this results in a building block size of 600 mm.

In this context, we had the idea to use a (large) *diffractive* lens instead of the achromat. For diffractive lenses, the Petzval curvature disappears [23, p. 27] and it should therefore be possible to obtain a smaller building block size with still acceptable imaging errors. Indeed, raytracing showed that even a 100 mm diffractive lens still offers enough precision. However, the overall performance had to be kept in mind. A more detailed analysis showed that the gain in alignment precision had to be paid by losses in throughput and uniformity: The efficiency of an achromat with anti-reflection coating is about 99 %, whereas the efficiency of a diffractive lens is typically in the region of 90 % in the center and 60 % in the corners [24]. The project was finally not realized.

d) Other aspects

When dimensioning such a configuration, one must not forget to control the *period* Λ of the grating. To obtain large deviation angles (grating close to LCTV), a small period is required. The system designer has to verify with the grating manufacturer that the feature size within the repeating structure does not get too small. Small angles (grating far from LCTV) correspond to a large periodicity. In this case, one has to verify that the illuminating beam covers enough periods (at least 3 to 4) in order to get sufficient diffraction performance.

The *size of the generated spots* is related to the *beam diameter*. A plane wave incident on a lens with focal length f and circular aperture of diameter D, generates a spot with a minimal diameter d, which is given by (cf. Appendix B.2)

$$d = 2.44 \frac{\lambda f}{D} \quad . \tag{2-4}$$

If the beam diameter is smaller than the lens aperture, we have to replace the lens aperture diameter D by the beam diameter D'. A smaller spot size requires a larger illumination beam (e.g. for a 300 mm lens the illuminating beam has to have a diameter ≥ 16.6 mm in order to obtain a spot size ≤ 21.5 µm). When increasing the beam diameter, one has to keep in mind the consequences of this to later stages of the system (e.g. crosstalk). The spot size can intentionally be chosen a bit too large. This results in less throughput and eventually some crosstalk, but it makes the setup less sensitive to aberration related problems.

Another point to consider is the *orientation of the lens:* An achromat is normally oriented with its flat side towards the divergent or convergent part of the beam and the curved side towards the collimated part of the beam, in order to reduce spherical aberrations [15, p. 235]. In the discussed setup, however, we have two contradictory interests: 1) each *individual beam* generated at the grating is collimated and will be focussed by the lens. 2) the *whole set of beams* is diverging and will be "collimated" by the lens. From the point of view of the *individual beams,* the achromat should thus be oriented with its flat side towards the input LCTV, whereas from the point of view of the *set of beams* it should be oriented with the flat side towards the laser. Raytracing proves what can be expected intuitively: with the flat side towards the input LCTV, the quality of the individual spots is better (smaller spot size), but the deformation of the whole image gets too big at the edges. The flat side towards the grating results in larger spot sizes in the corners of the array, but the overall image geometry is much better. With the raytrace software [22], we quantitatively compared the two cases by counting the total number of rays that correctly hit the corresponding pixel aperture. The orientation with the flat side towards the grating gave significantly better results.

In an early design phase, the use of focussing fan-out elements has been considered. In such a device, the lens and the grating are combined into one diffractive element [23]. Problems related to glass lenses (focal length variations, field curvature) can be avoided with this method. The drawback is the divergence of the generated set of beams. This can be immediately seen if we look at Fig. 2.9 and Fig. 2.10. In both cases, the focussing fan-out corresponds to the extreme case where the grating is moved to close contact with the lens. Due to the large divergence and the multi-stage nature of our setup, we did not consider this approach further.

e) Conclusion for the lens-grating-combination

Several possible arrangements of a diffraction grating and a lens have been discussed. In general, the choice for one arrangement has to be made in the context of the whole system (precision requirements, requirements of following system stages, system size restrictions, etc.). In our case, the best arrangement is an achromat of long focal length (≥ 300 mm) with its flat side oriented towards the grating and the grating positioned in the front focal plane of the lens. The influence of focal length errors and remaining aberrations can be reduced by a choosing an incident beam diameter that results in a spot size slightly larger than the pixel apertures.

f) Experimental details

We worked with a binary phase grating fabricated by Weible OpTech, Neuchâtel. Designed for a wavelength of 488 nm, the grating produces 16 x 16 diffraction orders. It is an "even-orders-missing" design, i.e. the orders 0, 2, 4, etc. are suppressed. The periodicity Λ of the grating is ~268 µm, which results in a pitch of 1092 µm in combination with a 300 mm achromat.

The measured efficiency of the grating was around 68 %. About 0.5 % of the incident light went into the (unwanted) zero order. The uniformity was < ±5 % (deviation from the mean value).

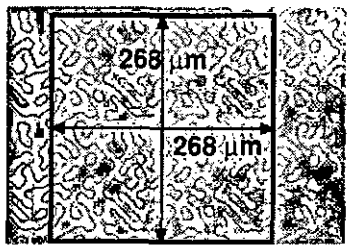


Fig. 2.11: Photo of the Dammann-grating. The structure produces 16 x 16 diffraction orders at 488 nm. The frame shows one unit cell of the repeating structure.

2.3.3 VCSEL-arreys

A third method to generate the desired input vector are vertical cavity surface emitting laser arrays (VCSEL-arrays). The interest in this technology has been rapidly growing in the last few years due to the potential usefulness of VCSELs e.g. in free-space optical interconnects and other applications. At the beginning of this project, we had the chance to test one of the first available 16 x 16 arrays of VCSELs. However, this element was clearly a prototype and suffered from many restrictions. The technology has matured in the meantime and a lot of these early problems can be considered as solved. To maintain generality, some general properties of state-of-the-art VCSELs will first be discussed before our experiences with the prototype are reported. Most of this information has been provided by Dr. Karlheinz Gulden, head of the VCSEL-group at CSEM Zürich.

a) Overview of the technology

As of today, there are a lot of *laboratory prototypes* around. The best arrays have threshold currents of ~0.6 mA, maximal output power of 4 mW and efficiencies of ~30 %. There is, however, only little literature about the simultaneous use of all the lasers in an array. Most people just test single lasers within an array.

Commercial VCSEL-arrays are only offered by very few manufacturers, e.g. CSEM Zürich. Typical array size is 8 x 8 VCSELs emitting at 960 nm. Typical parameters are: 2 mA threshold current, 1.5 mW optical power per VCSEL, 5–10 % efficiency, data rate > 1 GBit/s, divergence (FWHM) < 10°.

The maximum *optical power* per VCSEL is limited by the maximum tolerable heat dissipation on the array, which in turn depends on the heatsinking and the pitch of the VCSELs. Single VCSELs with > 100 mW optical power and 50 % efficiency have been demonstrated, but such VCSELs cannot be used simultaneously within an array. For the near future, arrays with 1–2 mW optical power and ≥ 20 % efficiency are a realistic expectation.

The *nominal emission wavelength* of VCSEL-arrays is typically 850 nm (telecommunication standard for short distance optical links). Another frequently used wavelength is 980 nm (the GaAs substrate is transparent at this wavelength, which is advantageous for integration technologies, e.g. flip chip technology). Within the next five years, VCSELs in the telecommunication window at 1.3 µm to 1.5 µm can also be expected. VCSELs in the range of 630 nm to 700 nm have been demonstrated in the laboratory. The commercialization of such prototypes will depend on the applications that demand a specific wavelength.

The *wavelength uniformity* over an array of VCSELs depends on the fabrication process and on the size of the array. Wavelength uniformity of ≤ 0.2 nm over an 8 x 8 array with 2 mm x 2 mm size has been demonstrated, for a 16 x 16 array, the limit can be expected around 0.5 nm. *Typical* wavelength uniformity for laboratory prototypes is rather in the order of 10–15 nm.

Wavelength variations during operation are due to thermal expansion of the VCSEL resonator cavity. They depend on the quality of the heat sink, on the efficiency of the lasers and on the number of simultaneously active lasers. In the extreme case, variations of a few nm are possible.

The *polarization uniformity* over an array of VCSELs is still a problem. In most of today's arrays, several polarizations can be observed across the array. Moreover, many of today's VCSEL arrays have multi-mode VCSELs. A stable polarization direction over a whole array has been demonstrated using suitable fabrication processes [25], but this is not yet a standard.

The physical size of VCSEL arrays is limited by the fabrication technology. Important in the context of this work is the fact that there is no technological problem to fabricate arrays with a large pitch (e.g. 1 mm) between the VCSELs. Generally, the price will increase with the dimensions of the array and one will have to use flip chip technology and bottom emitting VCSELs for large arrays.

A roadmap for the future development of VCSELs [26] predicts arrays of 64 x 64 single mode VCSELs with 0.4 mA threshold current, 0.5 mW optical power and 2 GBit/s bit rate per VCSEL for the year 2007. If optical techniques should prove to be successful in interconnect applications, a standardization of the VCSEL technology, reduced cost and increased availability can be expected.

b) Use of VCSELs in an optical system

Some points have to be kept in mind when the use of VCSELs is planned in an optical system.

Diffractive elements such as Fresnel lenses and gratings are *wavelength sensitive*. Planning to use such elements in combination with VCSELs, the system designer

has to check the influence of the expectable wavelength variations to the system. In a diffractive fan-out, for example, the pitch of the generated spots is proportional to the wavelength (cf. section 2.3.2). The focal length of a Fresnel lens, on the other hand, is inversely proportional to the wavelength (cf. Appendix C.1.1).

Currently, only VCSELs *emitting in the infrared* are commercially available. The spectral behavior of other planned system components (sensitivity, reflectivity, etc.) has to be kept in mind when considering different input sources. To work with a visible wavelength is certainly not a must, but a personal experience is that it is clearly easier to align and work with a complex system using visible light.

The *divergence* of the VCSEL beams influences the kind and dimension of the optical system that will relay the input image on the VCSEL-array to further system stages. The goal is to collect all the emitted light and to avoid crosstalk between channels. Assuming a full divergence angle of 10°, trigonometry yields a rough estimation for the beam diameter D at a distance z from the VCSEL: $D \approx 0.175\ z$. For a channel pitch of 250 µm these beams start to overlap at $z \approx 1.4$ mm, for a pitch of 1 mm at $z \approx 5.7$ mm.

Liquid crystal spatial light modulators are *polarization sensitive*. An array of VCSELs with a non-uniform (and normally not predictable) polarization distribution is not suited for use with liquid crystal devices.

A standard *pitch* of today's VCSEL-arrays is 250 µm. In combination with other array devices with different pitch, a magnifying optics is needed to match the different pitches, which can eventually introduce undesired aberrations to the system.

And finally, of course, the available *optical power* has to match the requirements of the light sensitive devices of the setup at the available VCSEL wavelength, e.g. in our case the sensitivity of the LCLV (cf. section 3.5.3).

c) Description of the tested prototype

The prototype that we could test was fabricated in 1994 at PSI Zürich (now CSEM Zürich) and had 16 x 16 individually addressable VCSELs. The pitch of the VCSELs was 280 µm x 230 µm. Most VCSELs emitted at 848.3 nm; towards one corner the wavelength dropped to ~846 nm. The threshold current was ~14 mA, the maximal optical output power for an individual VCSEL was ~1.1 mW, average was about 0.6 mW. The divergence of the beams was ~10° (FWHM).

Fig. 2.12.a shows a photograph of the 16 x 16 VCSEL array prototype. Visible are the bonding wires and the chip with some VCSELs (not lasing, but working as light emitting diodes; cf. below) that form the letters PSI, EPFL and IMT. Fig. 2.12.b shows the support with an improved Peltier-cooler (placed behind the VCSEL array), driver electronics and the EPROM-emulator.

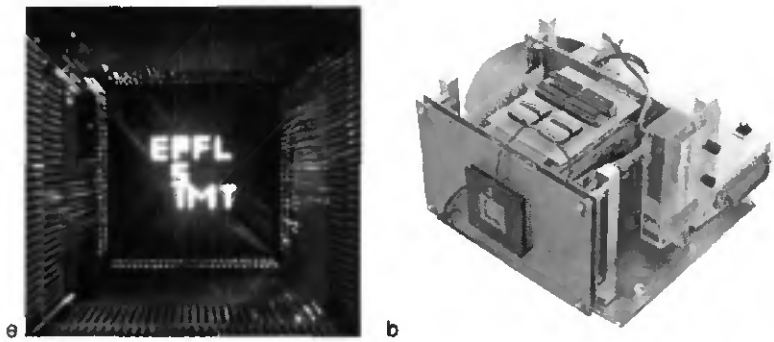


Fig. 2.12: a) Photograph of the 16 x 16 VCSEL array prototype that was built at PSI Zürich (now CSEM Zürich) and tested at IMT Neuchâtel. Visible are the bonding wires and the chip with some VCSELs (working as LED; cf. text) that form the letters PSI, EPFL and IMT. b) Photo of the support with improved Peltier-cooler behind the VCSEL array, driver electronics and EPROM-emulator.

Due to the large threshold current, low efficiency and non-ideal heatsinking, the array suffered from numerous heat related problems: With increasing heat, the emitted optical power normally decreased. Simultaneous use of more than 10 lasing VCSELs was not possible. However, even if the VCSELs were not lasing, they could still be used as light emitting diodes (LED, cf. Fig. 2.12.a). Due to the special construction of the support with the driver electronics, the Peltier-cooler had to be positioned besides the array, which was not very effective and introduced non-uniform cooling. Clear differences between VCSELs close to the cooler and far from the cooler could be observed. About 24 % of the VCSELs emitted s-polarized light, 67 % p-polarized and 9 % emitted more than one transverse mode. With increasing heat, some VCSELs changed the polarization direction. The prototype VCSEL-array had to be addressed by an EPROM-emulator, which was a very slow method.

We would like to point out that in spite of all these limitations, we were able to perform a lot of interesting and useful work with this device. Very often, the element was used as an array of light emitting diodes. A lot of the early building blocks of our system have been tested with this kind of input illumination.

2.3.4 *Conclusion for the choice of the input generation method*

For an arbitrary optical system, the choice of one input generation method depends on many system specific factors, which have been discussed above. The decisions specific to our system are presented and a summary of some points discussed in this section is given in Table 2.A.

Method / Property	Illuminated microlens-array	Fan-out with grating and lens	VCSEL-array (tested prototype)	VCSEL-array (today's state of the art)
Image distortion	excellent (depending on the quality of the microlens-array)	good, but only for lenses with long focal length (≥ 300 mm)	source: excellent; distortions due to necessary magnifying optics	excellent; dilute arrays can eliminate need for magnifying optics
Dilute array possible	yes	yes, requires long focal length	magnifying optics needed; not ideal	yes (not standard; standard pitch = 250 µm)
Spot size at LCTV	large, depending on focal length	small, depending on beam diameter	does not apply	does not apply
Overall divergence	depends on focal length, ≥ 3° for a short f (avoiding spot size problems)	depends on position of grating and focal length (e.g. ≥ 3°, cf. left)	10°	< 10°
Achievable optical power	depending on laser source	depending on laser source	generally bad; heat problems due to high threshold current	1.5 mW per VCSEL (8 x 8)
Threshold current	does not apply	does not apply	14 mA	2 mA
Losses	bad; losses due to expanded illumination beam, spot size, LCTV polarizer	better; losses due to efficiency of Dammann-grating, ev. spot size, LCTV polarizer	does not apply	does not apply
Wavelength	any	any	848 nm, infrared	(red -) infrared
Wavelength uniformity	does not apply (only one source)	does not apply (only one source)	≤ 2 nm (0.24 %)	down to 0.2 nm
Wavelength stability	good, depending on laser source	good, depending on laser source	bad due to low efficiency and related heat problems	potentially good, depending on efficiency, # of active VCSELs and heatsinking
Polarization uniformity over the array	no problem	no problem	bad	demonstrated but not yet standard; usually multi-mode
Polarization stability	no problem	no problem	bad; changes with heating up	should be good in absence of heat problems
Addressing speed	video rate, limited by LCTV	video rate, limited by LCTV	slow, limited by non-optimized device driver	≥ 1 GBit/s
Availability	good	good	early prototype	possible, but expensive due to non-standard specs
Overall building block size	medium	large	small	small

Table 2.A: Comparison of three concepts for the generation of the input vector. Because the tested VCSEL-array was a very early prototype, a fourth column with state-of-the-art VCSEL characteristics has been added.

The *illuminated microlens-array* was successfully used in the matrix-vector-multiplier tests. It is the easiest method with the best spot array geometry. However, when we started to implement the optical feedback, we realized that the losses due to the illumination and due to the large spot size were not acceptable.

The *final setup* uses the combination of *diffraction grating and lens*. This approach resulted in enough throughput for the optical feedback, but it almost doubled the overall system size. The sacrifice was acceptable because small size was not the main goal of our work and because the system still remained relatively compact.

At the moment of the decision, the *VCSEL technology* was not yet mature enough and could not be considered for the work on our setup. However, I believe that it will be the source of the future. The technology has made enormous progress and the availability of reliable devices is steadily increasing.

The ideal VCSEL-array for an application like ours should emit in the visible, ideally somewhere between 630 nm and 700 nm. This would allow for the use of a promising LCLV from Hamamatsu Photonics (PAL-SLM) that has its sensitivity maximum at 700 nm, and it would also facilitate the alignment of the system. To use the device in combination with polarization sensitive components such as LCTV and LCLV, the polarization should be uniform and constant for all the laser diodes of the array. The pitch of the laser diodes within the array should correspond to the channel pitch in order to avoid additional matching optics. The wavelength variation should be ≤ 0.1 % in order to allow the combination with diffractive optics.

2.4 Telescopes and relay optics

The generated input image has to be relayed to a next stage of the system (in our case to LCTV2 with the weight matrix). A large image field and a low tolerance for imaging errors (cf. section 2.3.2) make this task a difficult one. The following discussion of several kind of telescopes that can perform this task is split up in two fundamentally different groups: conventional glass optics (or macro-optics) and micro-optics.

The kind of subsystem that is normally used in the following discussion is called an *infinite conjugate, telecentric image relay system* [27].The distance between the object and the first lens is equal to the focal length $f1$ of the first lens. The distance between the second lens and the image is equal to the focal length $f2$ of the second lens. The distance between the two lenses is equal to $f1 + f2$. If $f1 = f2$, this is also called a 4f imaging system. Advantages of this configuration are collimated beams in the region between the lenses and perpendicular incidence of the principal rays in the image plane. This kind of configuration is related to the astronomical (Kepler) telescope, where object and image are at infinity [21]. For simplicity, the term *telescope* will be used throughout this section to describe such configurations.

2.4.1 Conventionel telescopes

An introductory remark has to be made at the beginning of this subsection: with enough effort and resources, it is of course possible to design and construct well corrected lens systems for a specific task using glass optics. Camera manufacturers, for example, obtain impressive results using sophisticated design and optimization software. Such systems include mostly custom-made lenses and eventually lenses with aspheric surfaces. Such an approach was not considered for this project. We therefore looked at configurations that could be realized with moderately priced standard lab material. Easy alignability of the system was also an important criterion.

During the development of this project, we looked at several telescopes. When working with the VCSEL-array, we needed a magnifying telescope; later, with the dilute input array (LCTV), a 1-to-1 telescope was used. The main problem during all this work was a mismatch between the imaged spot array and the pixel array of the second LCTV (weight matrix plane). This mismatch was mainly due to field curvature and incorrect magnification.

Of the numerous setups that were compared, only telescopes with long focal lengths ($\geq$ 300 mm) were able to provide the required accuracy. This obviously corresponds to the result of section 2.3.2 (fan-out with grating and lens).

Fig. 2.13.a shows an example for a raytraced telescope that clearly suffers from field curvature (magnifying telescope, used in combination with the VCSEL array described in section 2.3.3, f1 = 40 mm, f2 = 100 mm;). Fig. 2.13.b shows what happens if one tries to align an array of spots with such deformations to the pixel array of a LCTV: a correct match can only be achieved in the central region. This example is of course an extreme case, but it illustrates the problem very well.

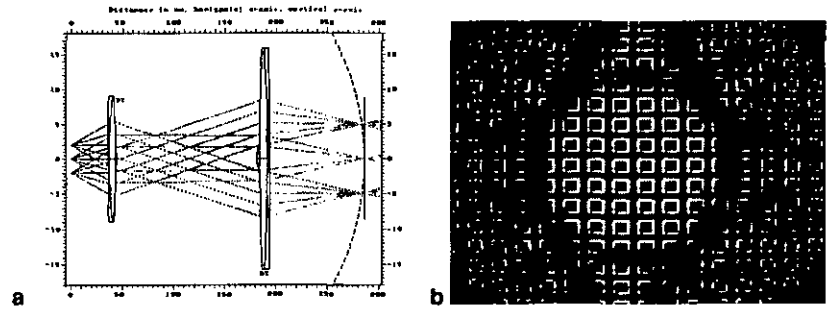


Fig. 2.13: a) Example of a raytraced telescope that suffers from field curvature (cf. text for parameters). b) For an image with such deformations, good alignment of the spot array with respect to the pixel array is only possible in the central region.

The problem of incorrect magnification comes from the fact that reasonably priced commercial glass lenses have (according to their datasheets) a tolerance for the focal length of about 1 % (as already mentioned in section 2.3.2). If the focal lengths of the two lenses are not identical, the resulting image will be magnified or reduced by a factor M, which is given by

$$M = f_2 / f_1 \quad . \tag{2-5}$$

In the worst case of $f_2 = 1.01 \cdot f_1$, we get an image which is 1.01 times larger than the object. In our case of an object size of 15 x 1092 µm = 16.38 mm, the image size would become 16.544 mm. The difference is 164 µm, which results in an 82 µm displacement of the spots at the edge of the array if the image is correctly aligned in the center. In order to remain within the acceptable 10 µm displacement, the difference between f_1 and f_2 must be smaller than 0.12 %. This is more than 8 times less than the tolerance given by the lens manufacturers.

As already mentioned in section 2.3.2, it is possible to overcome the problem of the focal length tolerance by using a pair of lenses instead of a single lens, which allows for a adjustable effective focal length. The alignment of such a system, however, is difficult because the position of front and back focal plane changes when the effective focal length is adjusted [21, section 2.10]. This is acceptable for a linear system [11], but becomes a nightmare in a setup like ours where the back end of the telescope is shared by two front ends (initialization and feedback loop).

Another option is of course to exactly measure the focal length of several lenses and work with a matching pair.

2.4.2 Micro-channel telescopes

An alternative to conventional telescopes for discrete images (like spot arrays) are microlens array telescopes. In the literature, they are often referred to as micro-channel telescopes [28-30]. In such a configuration, each channel has its own little telescope and each source (spot of the input array) is in a paraxial position. The advantages of this approach are obvious: Aberration problems related to a large field of view are completely eliminated. The overall quality of the image depends on the quality of the microlens array, which can be excellent (cf. Appendix C). Each individual microlens may of course still have aberrations, but their influence is only local.

Drawbacks of this approach are increased losses (depending on the kind of microlens technology; e.g. diffraction losses), and, as of today, cost and availability. Recent progress in replication technology [31] (cf. also Appendix C.1.3), however, will enable mass production of microlens arrays and thus lower the prices.

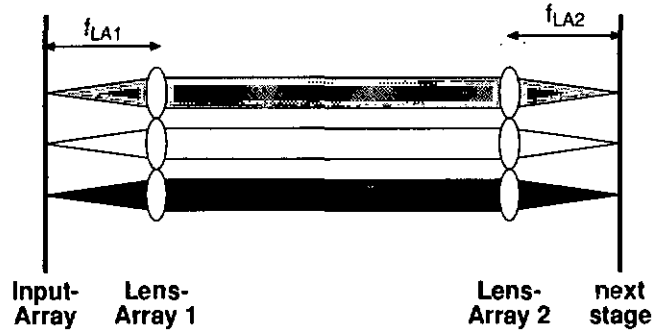


Fig. 2.14: An alternative to conventional telescopes for discrete images are microchannel-telescopes. Each source is in the paraxial position of its own little telescope. Aberrations related to a large field of view are eliminated and eventual remaining effects are only local. The telescope length is diffraction-limited.

One is tempted to arrange the microlens arrays in a 4f or telecentric configuration (distance equal to the sum of the focal lengths). This is an advantageous configuration for macro-optics, because it leads to perpendicular incidence of the principal rays in the image plane, but it is not necessary in a symmetric (1-to-1) microchannel-telescope, because every point of the source array is in the paraxial position of its corresponding microlens. Larger distances than *f1 + f2* between the two microlens arrays allow for larger telescopes which can be important when other components (e.g. a beamsplitter) have to be placed within the telescope or when the signal has to be relayed over a large distance (as in a feedback loop). Just using longer focal lengths in order to increase the telescope length is not a good solution, because this also increases the diffraction limited spot size (cf. section 2.3.1 and appendix B.2).

There is, however, a limitation to the maximal length of a microchannel telescope, which is again caused by diffraction: light beams cannot propagate over an arbitrary distance without increasing in width. A rough analysis [29] yields a maximal length L between the microlens arrays:

$$L \leq \frac{D^2}{\lambda} \quad , \tag{2-6}$$

where D is the channel pitch. This formula does not include the focal length of the used microlenses. In practice, however, lenses with a shorter focal length will lead to a shorter maximal length L, which can be understood if one looks at the properties of a Gaussian beam (cf. Appendix B.4). As a consequence, a compromise has to be found between a short focal length which minimizes the spot size and a long focal length which allows for longer telescopes (cf. also Fig. 2.7.b).

A remark about the *magnification* of the microchannel-telescope. Because we are dealing with an array of telescopes, there are two kind of magnifications involved.

The *channel-magnification* can be defined as the ratio between image spot size and object spot size. It is given by the ratio of the focal lengths of the two arrays. The *array-magnification* can be defined as the ratio between image array size and object array size. It is determined by the pitches of the object array and of the two microlens arrays (cf. Fig. 2.15).

The microchannel approach is therefore not only suited for symmetric 1-to-1 telescopes. Especially the computer generated diffractive microlens arrays allow a high degree of customization. Arrays of microlenses with locally variant properties have been demonstrated by Hessler *et al.* for a confocal microscopy application [32, 33]. But also refractive microlens arrays with a pitch different to the source pitch can do the task [34] (Fig. 2.15).

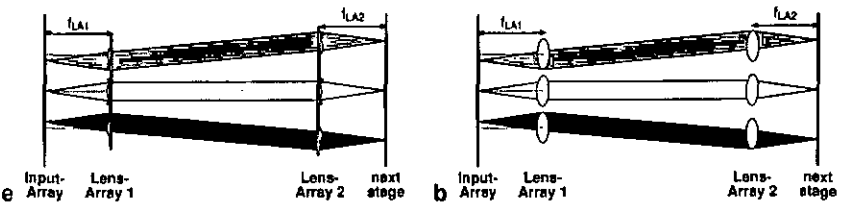


Fig. 2.15: Two related methods to implement a magnifying microchannel-telescope: a) Diffractive microlens arrays with locally variant properties or b) refractive microlens arrays with a pitch different to the input pitch can be used to adapt the pitch of an input array to the pitch of a following stage.

If it is possible, however, a symmetric setup without array-magnification should be preferred. The problems of a magnifying microchannel-telescope are related to the angle of incidence of the chief rays and thus to the overall divergence of the set of beams. If the second microlens array is moved along the z-axis, the position of their focal spots remains constant but the angle of incidence changes. The same happens if an offset is introduced to the collimated beams, e.g. by placing a beamsplitter between the two microlens arrays. All this is not the case if the collimated beams between the two microlens arrays are parallel (symmetric setup). As a consequence of this, the symmetric setup is more tolerant to alignment errors. This is a main advantage, especially also for the implementation of a feedback-loop, where the beams of the initialization have to match the beams of the feedback.

Moreover, in the array-magnifying configuration, the source is not in a paraxial position, which increases the local aberrations of every channel.

2.4.3 Conclusion: Choice of the telescope principle

We conclude that microchannel-telescopes are the better alternative for applications with a large number of discrete parallel channels. The classical 4f configuration might not be the best arrangement if every source spot is in a paraxial position. The limitations due to diffraction (spot size and beam spread) have to be considered for the choice of the focal lengths and of the telescope length.

2.4.4 Experimental comparison

During the early comparison of the two fan-out methods (cf. section 2.5), the number of channels has been limited to 5 x 5, mainly due to imaging errors. In the final setup we used microchannel-telescopes instead of bulk lenses, which effectively eliminated the problem of pattern mismatch. Fig. 2.16.a repeats Fig. 2.13.b, showing an image of a 8 x 8 spot pattern (frame), fanned-out to LCTV2 using conventional fan-out (cf. section 2.5.1) and bulk lenses. Only the spots in the central region match their corresponding LCTV-pixel. Fig. 2.16.b shows a 16 x 16 spot array fanned-out to LCTV2 using interlaced fan-out (cf. section 2.5.2) and a microchannel-telescope. On LCTV2, a set of alignment patterns (frame with diagonals) is displayed. By slightly displacing LCTV2 (cf. Appendix D), we could verify that each of the 65'536 beams correctly matches its corresponding pixel.

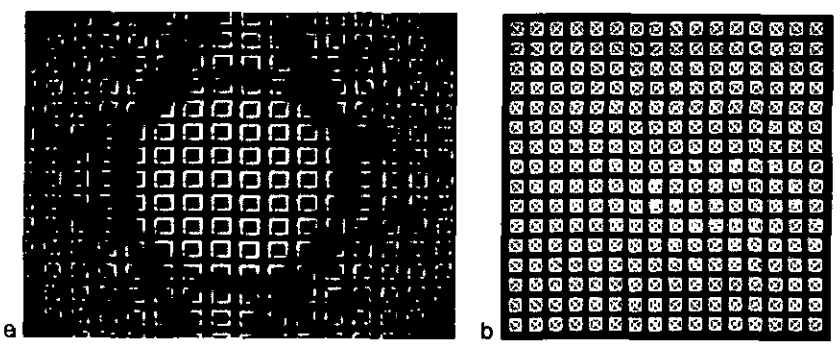


Fig. 2.16: a) Mismatch between a spot array and the pixel array of LCTV2 caused by field curvature using macro-optics; b) Correct match of each of the 65'536 spots to its corresponding pixel using the microchannel approach. (Remaining inhomogeneity in this picture is due to vignetting of the image acquisition system).

Most of the microlens arrays used for these (and other) tests are continuous-relief diffractive Fresnel elements and they were fabricated in photoresist with the laser beam writing system of the CSEM Zürich (former PSI Zürich) [35, 36]. The main series of experiments was made with various arrays of 16 x 16 microlenses with 11.54 mm, 50.00 mm and 80.00 mm focal length respectively (at a design wavelength of 488 nm) and a pitch of 1092 µm x 1092 µm (corresponding to 26 pixel pitches of 42 µm x 42 µm of our InFocus LitePro 850 LCTVs). The lateral size of the lenses was equal to the pitch, the fill factor thus 100 %. Additional details about these and other tested microlens arrays are provided in Appendix C.

2.5 The matrix-vector-multiplier: fan-out and fan-in

In neural network applications, but also in switching tasks, it is often necessary to distribute the information of one channel to other channels. This can be achieved optically by a fan-out and a subsequent fan-in. The conventional method is to reproduce the whole input image N x N times and then feed each of the generated sub-images into one channel of the output image. An alternative method (that to our knowledge has been demonstrated in this work for the first time), is to replicate each channel individually [37]. In order to link each output channel with every input channel, the output image has in this case to be a superposition of all the generated sub-images. Fig. 2.17 illustrates the two methods.

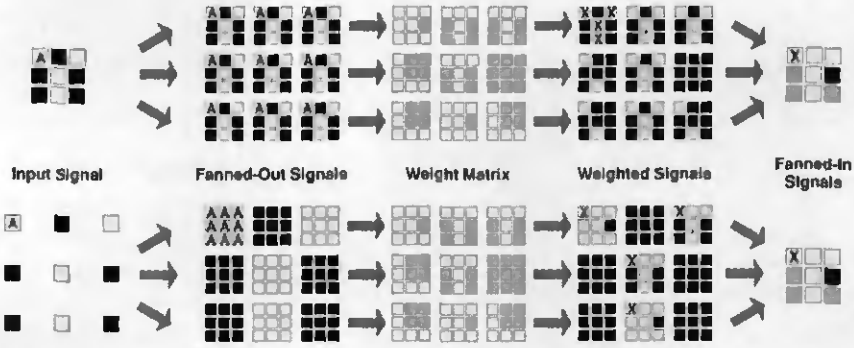


Fig. 2.17: Two fan-out concepts: The conventional method (top) is to replicate the whole image. For the interlaced fan-out (bottom), each channel is replicated individually. In order to obtain full connectivity, the fan-in also has to be different.

From a mathematical point of view, both operations correspond to the multiplication of an input vector (represented as a 2D–image) with a weight matrix (represented as

an array of sub-images). This operation is (together with the non-linear threshold function of the neurons) a central element of the Hopfield-model. The matrix-vector-multiplier can therefore be considered as the heart of the optical setup.

The two fan-out / fan-in concepts will be presented and compared in the following.

2.5.1 Conventional fan-out and fan-in

e) Fan-out

The conventional methods of replicating an input image are closely related to the points discussed in sections 2.3.1 and 2.3.2 (input generation), where the signal of a single source was fanned-out and modulated. The same considerations as there apply also for the case of a whole image that is fanned-out and has to be modulated.

Again, we thus have the method of an illuminated microlens array and of a fan-out with a diffraction grating. Examples for both methods are illustrated in Fig. 2.18. Advantages and disadvantages of both methods are the same as described in sections 2.3.1 and 2.3.2.

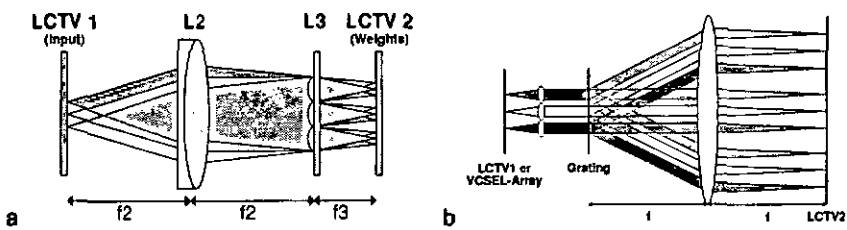


Fig. 2.18: Two possible implementations for the replication of the complete input array (conventional fan-out): a) using a microlens array, b) using a grating

b) Fan-in

After modulation of the fanned-out signals (e.g. by LCTV2), all the pixels of one sub-array have to be fanned-in to one output channel.

A *true optical fan-in* should create a superposition of all the spots of a sub-array. This can be achieved by combining a microlens array of small pitch (corresponding to the pixel pitch) together with a microlens array of large pitch (corresponding to the channel pitch) (Fig. 2.19.a). The focal length of the first microlens array will be rather short in order to avoid a large f-number (diameter of each microlens ≤ pixel pitch). Ideally, f4 should be chosen such that LA4 can be glued directly onto the LCTV in order to reduce alignment problems.

In hybrid systems (i.e. electronic feedback), a pseudo fan-in is mostly used. The weight matrix output is imaged to a detector array where the actual summation is performed by the detector cells (Fig. 2.19.b).

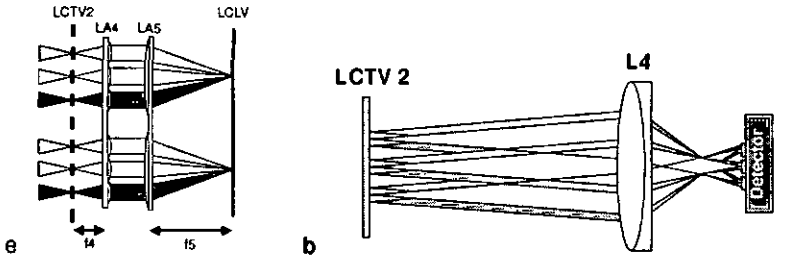


Fig. 2.19: In the conventional case, a true optical fan-in requires two microlens arrays of different pitch (A). Other methods use a pseudo fan-in with optical reduction of the sub-arrays and electronic summation by a detector array (B).

c) Experimental demonstration

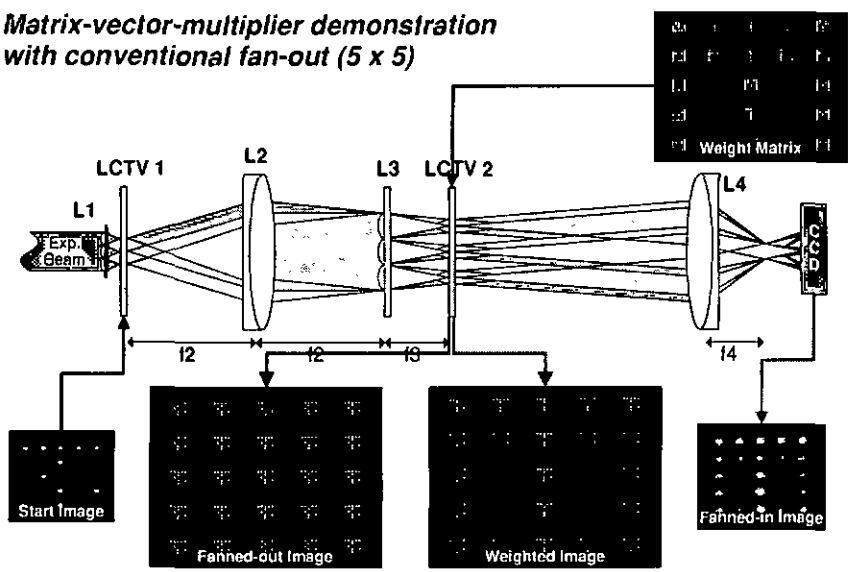


Fig. 2.20: Experimental demonstration of a 5 x 5 matrix-vector-multiplier with conventional fan-out and pseudo fan-in.

Fig. 2.20 shows an experimental demonstration of a matrix-vector-multiplier using conventional fan-out. The number of channels was limited to 5 x 5, mainly due to field curvature (cf. Fig. 2.13.b). The output of the system is only reduced and slightly defocalized in order to produce a pseudo fan-in. The algorithm that has been used to calculate the weight matrix is explained in section 4.1.

2.5.2 Interleced fan-out and fen-In

e) **Fen-out**

The pitch of the input array has to be large enough in order to allow for a replication of each input signal within its corresponding channel (dilute or wide spaced array). Of course it is possible to magnify a narrow spaced input array, but this can lead to imaging errors, magnification errors, etc. When the pitch of a spot array is magnified, we have to keep in mind that the size of the spots is normally magnified as well (except in the case of space variant arrays, cf. section 2.4.2). The ratio between pitch and spot size of the input array has therefore to be large enough in any case.

Whenever possible, we recommend to use a dilute input array with a pitch equal to the channel pitch. This allows for a constant channel pitch throughout the system and perpendicular incidence of the principal rays. Further advantages of the symmetric situation have been discussed in section 2.4.2.

The replication of each channel follows the same principles as all the other fan-outs that have been discussed so far. We can provide an array of microlenses for every channel (standard problem: uniformity of illumination) or we can perform the fan-out with a grating and a lens per channel. Both concepts are illustrated in Fig. 2.21. Due to the analogy to above, we do not want to repeat the details here. The configuration with grating and microlens array (Fig. 2.21.b) has been used in the final setup. Another possible implementation is shown below with the experimental results.

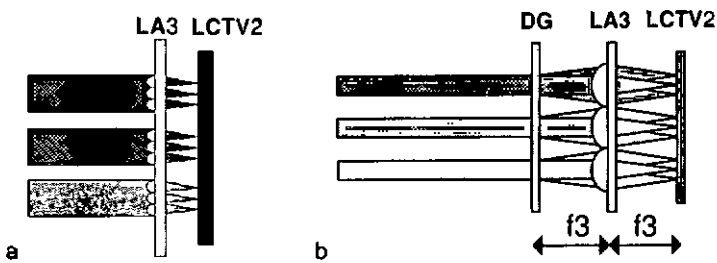


Fig. 2.21: The replication of each individual channel (interlaced fan-out) can be made using the same concepts as for the other fan-outs that have been discussed so far. Advantages and problems apply respectively.

The position of the grating with respect to the microlens array also follows the same rules as in section 2.3.2 (input generation with a grating and a lens). Fig. 2.22 shows a drawing from the design phase that illustrates the influence of the grating position (cf. also section 2.6.2 for a further discussion of this drawing). The ideal position of the grating is again the front focal plane of the microlens array.

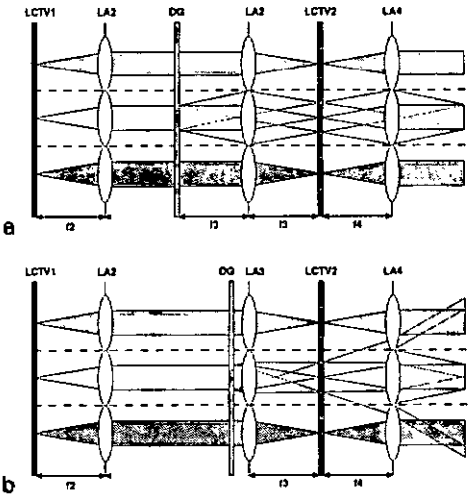


Fig. 2.22: Interlaced fan-out with grating: The ideal position of the grating DG is in the front focal plane of the microlens array LA3 (a). A position farther away from LA3 introduces crosstalk at LA3, a position closer to LA3 introduces crosstalk at LA4 (b).

b) Fan-In

Fig. 2.23 shows a schematic sketch of the complete matrix-vector-multiplier. In order to connect every output channel with every input channel, the output array has to be a superposition of all the sub-arrays that have been generated by the fan-out. This can be achieved in a very elegant way by combining a microlens array with a lens. The microlens array has the same pitch as the sub-arrays (i.e. the channel pitch). The distance between LCTV2 and the microlens array is equal to the focal length *f4* of the microlens array. Every microlens can be considered as a Fourier transformer that transforms spatial information into angular information. Light emerging from corresponding spots of the sub-arrays of LCTV2 (e.g. from the upper left spot of every sub-array, cf. Fig. 2.23 and Fig. 2.17) travels in the same angular direction after passage through the microlens array. L5 performs the re-transformation from angular to spatial information, which results in a superposition of the corresponding spots in the back focal plane of L5 (output plane).

The distance between the microlens array LA4 and the lens L5 only influences the angle of incidence of the beams on the LCLV. If this distance is equal to f4 + f5, we

obtain perpendicular incidence of the principal rays. Because the LCLV is the last stage of the system, this is not so important this time. On the contrary, our experiments showed, that it is best to have L5 at a distance f4 from LA4 because of the large divergence of the set of beams emerging from LA4. The spacing of the spots in the output plane is determined by the ratio of f5 to f4 and does not change with the position of L5.

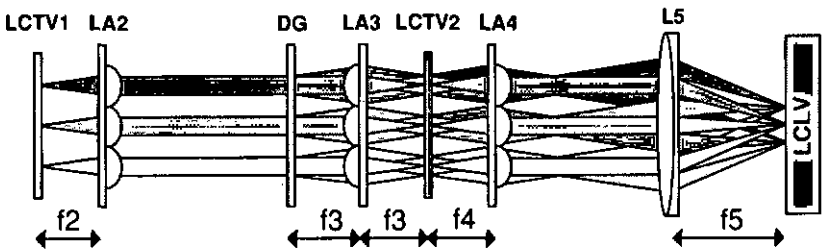


Fig. 2.23: Possible realization of an optical matrix-vector-multiplier with interlaced fan-out and subsequent fan-in.

c) Experimental demonstration

Fig. 2.24 shows an experimental demonstration of a matrix-vector-multiplier with interlaced fan-out. This demonstration was made with the same hardware as the demonstration of the conventional fan-out (cf. section 2.5.1).

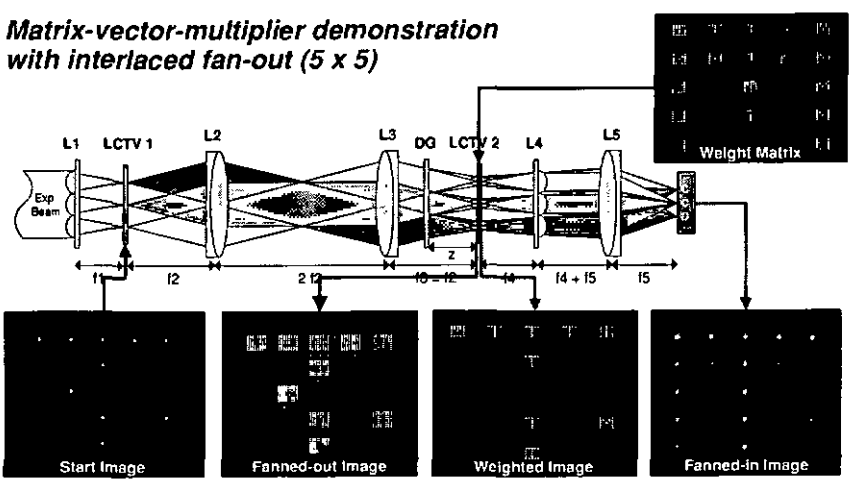


Fig. 2.24: Experimental demonstration of a 5 x 5 matrix-vector-multiplier with interlaced fan-out.

2.5.3 Conclusion: Choice of the fan-out principle

Interlaced and conventional fan-out are just inverse versions of the same thing. Fig. 2.25 shows another schematic representation of the two fan-out principles. In the conventional case, the signal is distributed to the other channels before the weight matrix and remains within one channel after the weight matrix. In the interlaced case, the signal remains within one channel before the weight matrix and is distributed to the other channels after the weight matrix. What makes now the difference between the two methods?

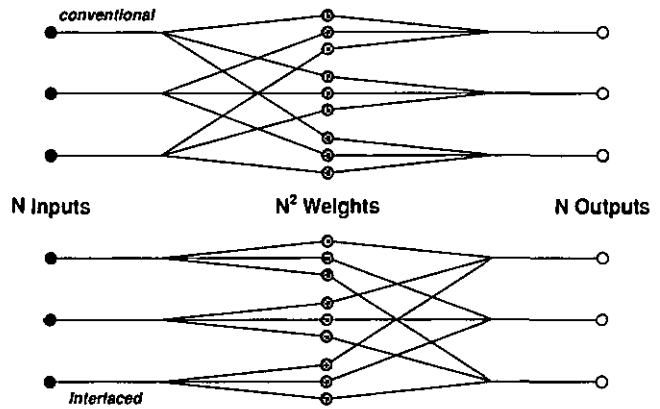


Fig. 2.25: Schematic representation of conventional (above) and interlaced (below) fan-out. Each method is the inverse of the other method.

Optically, the distribution of the signal to other channels is related to large angles and a large field of view. A large field of view, however, increases the importance of aberration related problems, as we have seen in the previous sections. The question is thus: is it better to have these problems before or after the weight matrix? If we compare the two tasks of matching N^2 spots to the weight matrix LCTV (a pixelated device) and of matching N spots to the LCLV (which in our case is not pixelated), we clearly prefer to optimize the situation for the first task. We therefore claim that the interlaced fan-out principle is better suited for the optical implementation of a full interconnect with a large field of view.

Moreover, the interlaced fan-out is ideally suited for an implementation with a symmetric microchannel-telescope (whose advantages have been discussed in section 2.4.2). The smaller angles of the interlaced fan-out result in a larger minimal feature size of the grating required for the fan-out, which can be an advantage for the fabrication of the grating.

Conventional and interlaced fan-out are inverse configurations, as mentioned above. This can also be seen in the optical implementation: The fan-in of the interlaced

method (lens-array and lens) corresponds to the inverse of one of the possible fan-outs of the conventional method. However, not all of the optical concepts are reversible: $N \times N$ beams incident on a grating, for example, will normally not unify to one beam (unless the phases of the incident beams correctly match). As a consequence, the fan-in of the conventional method cannot be as elegant as the fan-out of the interlaced method. We believe that the simple fan-in of the interlaced method (requiring the alignment of only one microlens array) is another plus for this method.

2.6 Complete system with optical feedback loop

In this section, we will combine the building blocks discussed in the previous section to a complete system. The compromises that had to be made between contradictory requirements will be described. For the better understanding, we repeat here the figure with the complete optical setup.

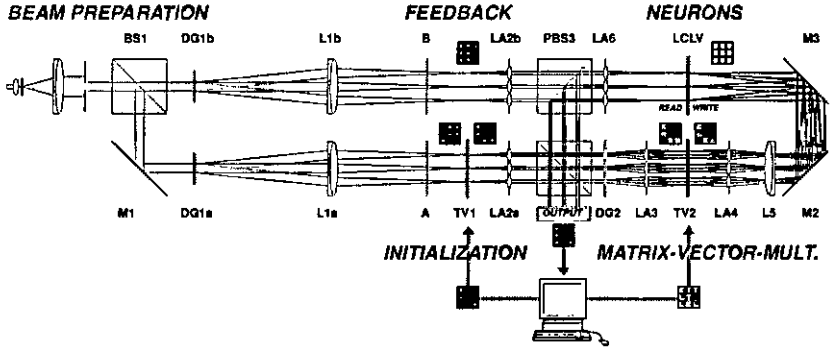


Fig. 2.26: Complete optical setup

2.6.1 Spot size, telescope length and opto-mechanics

To determine the absolute focal lengths of the telescope LA2–LA3, we had to find a compromise between crosstalk and the spot size at LCTV2 (small focal length desirable) and a reasonable telescope length, which offers enough space for the beamsplitter and the Dammann-grating (long focal length desirable). Initially we

decided for an afocal configuration with f2 = 50 mm. Later, we realized that it is better to use a shorter f2 (11.5 mm, value chosen for availability reasons) in order to avoid crosstalk (cf. Fig. 2.5). Dropping the unnecessary afocal arrangement allowed for a still long enough telescope. f3 = 11.5 mm was determined by the available space and by the value of f4 (see below).

To illustrate the opto-mechanical constraints, Fig. 2.27 shows a detail of the setup with BS2, DG2, LA3, LCTV2, LA4, L5 and M2. The elements around LCTV2 lie within a distance ≤ 3.5 cm and have all the necessary degrees of freedom for the alignment (normally x, y, z and ϑ_z). The independence of these degrees of freedom massively facilitates the alignment of the components.

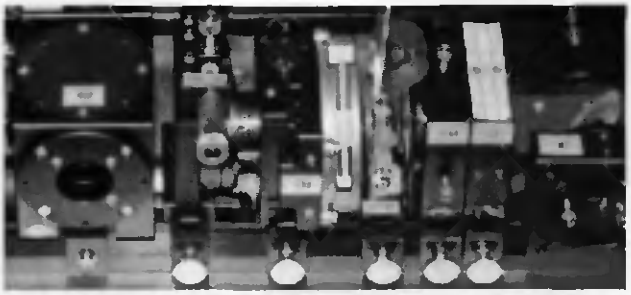


Fig. 2.27: Detail of the setup. (from left) BS2, DG2, LA3, LCTV2, LA4, L5, M2. The elements around LCTV2 lie all within a distance of ≤ 3.5.mm and have all the necessary degrees of freedom for the alignment.

2.6.2 Local beam diameter and performance of Dammann-grating

The diameter of the collimated beams within the telescope LA2–LA3 is determined by the diameter of the beam illuminating the first Dammann-grating and by f2. Here, we had to find a compromise between low crosstalk at LA3 (small beam diameter desirable) and performance of the Dammann-grating DG2 (large beam diameter desirable, covering ≥ 3 periods). Fig. 2.28 shows an illustration of the problem.

Because of the small diffraction angles of the interlaced fan-out, the grating has a rather large period of 268 µm. As a compromise between crosstalk and grating performance, we decided for a beam diameter of ~0.7 mm (covering 2.6 periods of the grating).

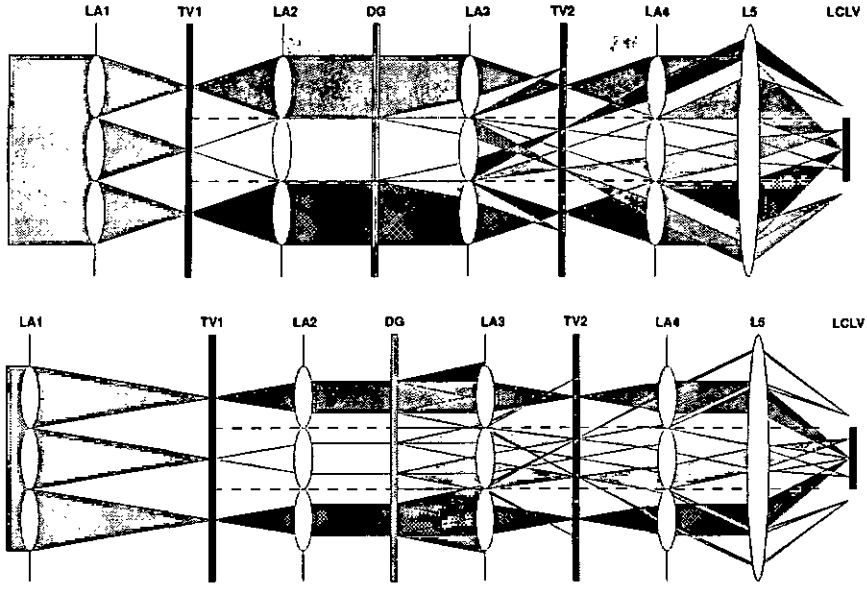


Fig. 2.28: Local beam diameter at the grating

A remark about the crosstalk at LA3. In Fig. 2.28, we can see that the light which goes into another channel at LA3 will finally not reach the LCLV. Therefore, the result is rather a loss of light than real crosstalk. However, we should not tolerate losses either, because they introduce a systematic error in the system (these losses are more severe for higher diffraction orders). Neural networks are in principle able to accommodate such imperfections in the weight matrix. However, this compensation alone will already use up a good part of the available contrast of the spatial light modulator. Losses should therefore be avoided as much as crosstalk.

Initially, we worked with an illuminated microlens array at the input (cf. Fig. 2.28). In order to obtain the desired beam diameter of ~0.7 mm, the focal length of this array was chosen to be 80 mm. Because of the spot size related problems of this configuration (cf. section 2.3.1), we later used a fan-out with Dammann-grating and lens, as shown in Fig. 2.26. The diameter of the collimated beam that illuminates the grating was chosen to be 18 mm in order to provide the desired beam diameter in combination with LA2. This value corresponds roughly to the value required to get a small enough spot size at LCTV1 (cf. section 2.3.2). Fig. 2.29 illustrates the relations between the different parameters. In practice, we often used a slightly smaller beam, which results in spot sizes that are a bit larger than the pixel apertures. This reduced the influence of the remaining field curvature of L1 and made the input module less sensitive. The resulting loss in intensity and the crosstalk introduced by diffraction at the LCTV1 pixels remained still acceptable.

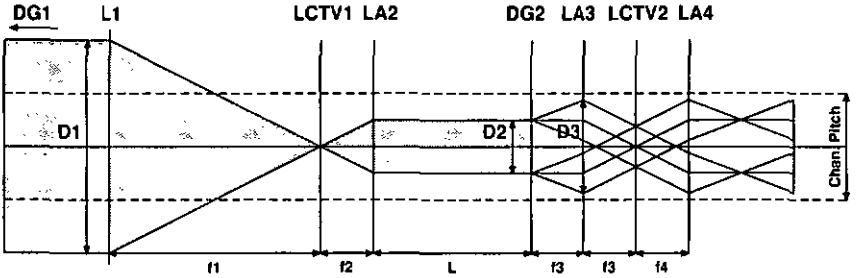


Fig. 2.29: Scheme of the light propagation within one channel of the matrix-vector multiplier. If the diameter D1 of the incoming beam is increased, the spot size in the plane of LCTV1 decreases, which reduces the effect of diffraction at the LCTV1 pixels. On the other hand, the overall diameter D3 at LA3 and LA4 increases, which results in losses (cf. text). A smaller diameter D2 reduces the performance of the grating DG2.

2.6.3 Matching the read-out signal to the write signal

In the feedback mode, the signal on the write side of the LCLV has to match the signal on the read side. The magnification of LA4–L5 has therefore to correspond to the pitch of the channels in number of pixels (26 in our case). For symmetry reasons f4 = f3. This results in the most effective use of the available lens aperture. f3 should be as small as possible to allow for a smaller period of the Dammann-grating, which (for a given beam diameter at the grating) increases its performance. However, f3 cannot be arbitrarily small for mechanical reasons (holder construction and alignment) and because of the limited feature size of the Fresnel-type microlens arrays. We have selected f5 = 300 mm (commercial achromat) and thus f3 = f4 = 11.5 mm.

The magnification of LA4–L5 results in an output array with a "spot" size of 26 times the pixel size (26 · 21.5 µm x 26 · 31.5 µm = 559 µm x 819 µm). This large spot size on the write side of the LCLV facilitates the matching of the write array with respect to the read array.

2.6.4 Feedback loop optics

In the first version of the system design [38], the microlens array telescopes were arranged in 4f–configurations, whenever possible. After our first experiences with the microchannel-telescopes, we realized, that this is not always the best solution (cf. section 2.4.2). As a consequence, the two initially planned telescopes between the

beamsplitters (PBS and BS2) have been omitted, which facilitated the alignment of the feedback loop. To avoid problems with beam spreading (cf. section 2.4.2), the focal length of LA6 was chosen to be 80 mm.

2.6.5 Matching the polarization states of feedback and initialization

One last point to consider was that the *polarization* of the feedback beam matches the polarization of the initialization beam. The light emitted from the laser is s-polarized. In the OFF-state, the first LCTV turns the polarization by 90° and absorbs the p-polarized output with its integrated polarizer. In the ON-state the output remains s-polarized. The output of the feedback-beam has thus to be s–polarized too. To achieve this, the polarization of the laser beam is turned by 90° by a half-wave plate. The p-polarized light passes straight through the polarizing beamsplitter. In its OFF-state, the LCLV does not turn the polarization and the reflected beam passes straight back through the polarizing beamsplitter and is lost. In the ON-state, the LCLV acts like a quarter-wave plate with mirror (cf. section 3.3) and the s-polarized output is deviated by the beamsplitter and correctly fed back into the system. The output of LCTV 2 is again s-polarized, but this is less important because the LCLV input is not polarization sensitive. Fig. 2.30 shows a schematic sketch of the polarization states of the ON-signals in the completed setup.

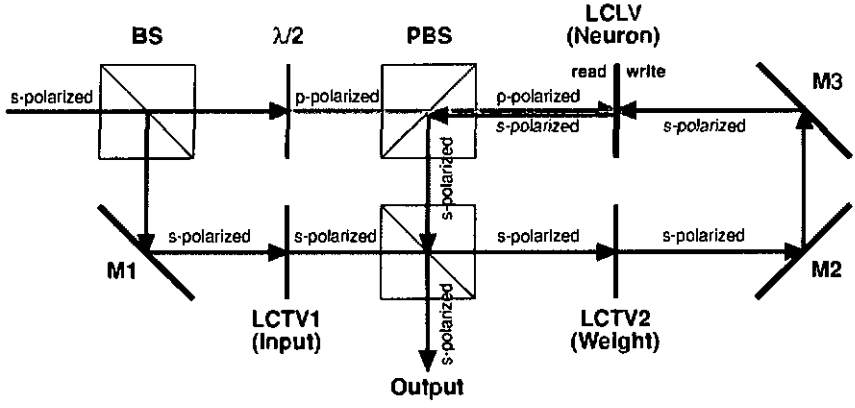


Fig. 2.30: Polarization states of the ON-signals in the completed setup.

2.6.6 Summery: Absolute values of system peremeters

Peremeter	Value	Specifications / Remerks
Wavelength	488 nm	Ar⁺-Laser, Spectra Physics, Mod. 165-09
LCTV pixel pitch	42 µm x 42 µm	from InFocus LitePro 850
LCTV pixel size	21.5 µm x 31.5 µm	from InFocus LitePro 850
Channels	16 x 16	
Channel pitch	1092 µm x 1092 µm (26 x 26 pixels)	maintained throughout system
f1 (a + b)	300.08 mm	Spindler & Hoyer Achromat, Ø 50 mm, @ 488 nm
f2 (a + b)	11.54 mm (16 x 16)	continuous relief Fresnel lens array, PSI-Z
f3	11.54 mm (16 x 16)	4-level Fresnel lens array, Weible OpTech
f4	11.54 mm (16 x 16)	= f5/26, 4-level Fresnel lens array, Weible OpTech
f5	300.08 mm	Spindler & Hoyer Achromat, Ø 50 mm, @ 488 nm
f6	80 mm (16 x 16)	4-level Fresnel lens arry, Weible OpTech
Period of gratings	268.2 µm	to obtain a pitch of 42 µm in combination with L3 and to obtain a pitch of 1092 µm in combination with L1

Table 2.2: Summary of system parameters.

2.6.7 Data acquisition and control software

In any mode, the state of the system can be monitored at the output of beamsplitter BS2. We mostly used a CCD-camera for the qualitative measurements. For further processing of the output (e.g. during in-situ training of the neural network), it is desirable to have separate signals for each channel. This can be achieved by post-processing the image provided by the frame-grabber that reads out the CCD-camera (slow) or by replacing the CCD-camera by a detector-array that directly provides a separate signal for every channel. The detector array was provided by CSEM Zürich (former PSI Zürich), the supporting electronics by Logitech SA.

Because of the matrix-nature of our data, we chose MATLAB™ as the programming platform for the acquisition- and control-software. Two data acquisition boards, a GPIB board and a frame grabber board allow to monitor and control all elements of the setup. This arrangement offers great flexibility: We can for example control the LCLV drive voltage and thus the feedback-loop gain (cf. section 3.5) in function of the system state and behavior.

The control software is soft-coded, i.e. most of the system parameters are stored in one central preference file. This allows for an easy adaptation of the code for a different configuration. Additional functionality can be added thanks to the modular

structure. For example, a feature to map out bad channels can be added in the future to reduce the influence of the LCLV non-uniformity (cf. sections 3.8 and 4.2).

2.7 Optical characterization

Fig. 2.31 shows a photo of the final version of the optical setup. On the left, DG1 and L1 are cut off to provide enough detail for the central part of the system. Some optical properties of this setup (without the LCLV) will be described in the following.

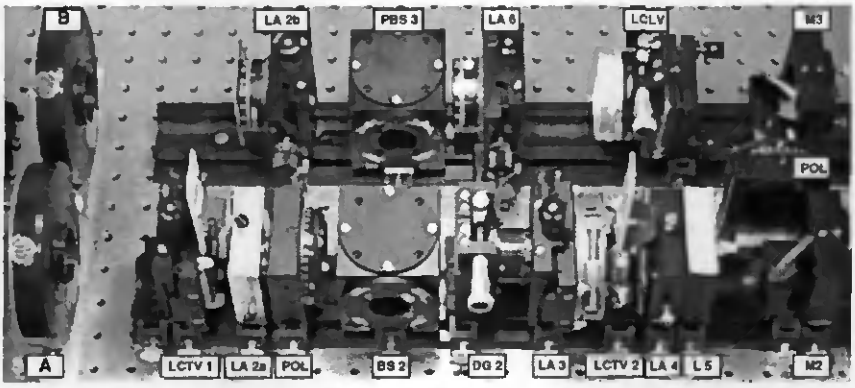


Fig. 2.31: Photo of the final version of the setup. The distance from shutter A (left) to mirror M2 (right) is ~55 cm.

2.7.1 Imaging quality

The image quality satisfies the demanding requirements throughout the system. Fig. 2.32.a shows a 16 x 16 input pattern emerging from LCTV1. Fig. 2.32.b shows a detail of the fanned-out signal before passage through LCTV2. 6 x 6 channels are shown, each consisting of 16 x 16 spots. Fig. 2.32.c shows the fanned-out signal after passage through LCTV2, which modulates a set of alignment frames to the signal. The fanned-in image at the output of the matrix-vector-multiplier is shown in Fig. 2.32.d. Imaging errors were reduced to a negligible level thanks to the microchannel-telescopes. The inhomogeneities in the corners of these pictures are due to the image acquisition system (vignetting).

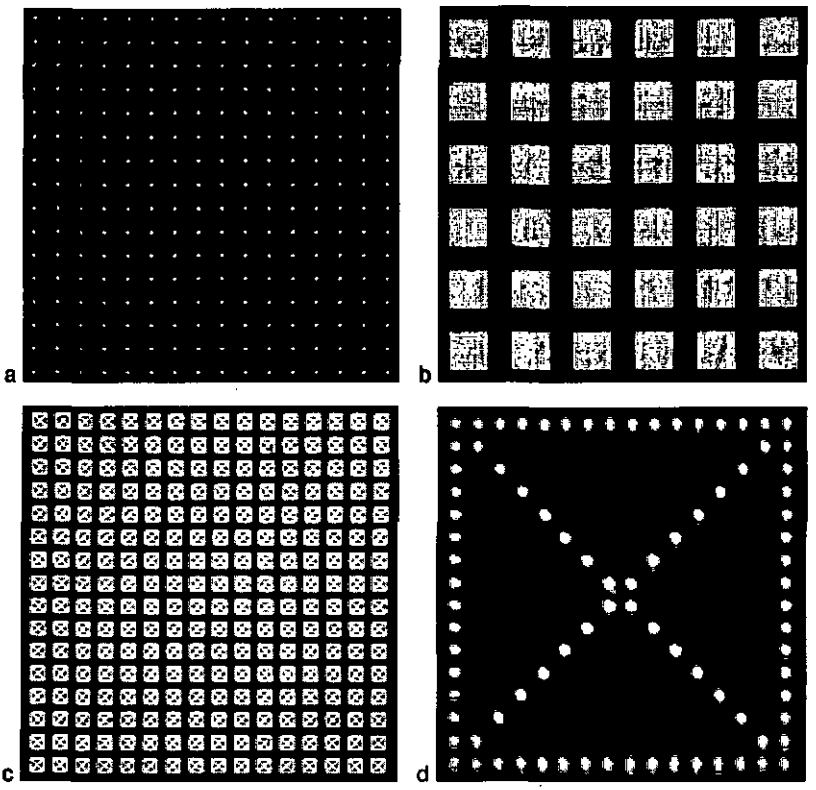


Fig. 2.32: Demonstration of the imeging quality of the setup: a) 16 x 16 spot erray emerging from the input LCTV1; b) Each channel is replicated 16 x 16 times; shown is a detail of 6 x 6 channels incident on LCTV2, each consisting of 16 x 16 spots; c) After passage through LCTV2, a test-pattern is modulated to the signals. Each of the 65'536 beams passes through its corresponding pixel; d) fan-in of the test-patterns (superposition of all sub-arrays)

It was possible to match the signal reflected back from the LCLV read side onto the pixels of LCTV2 with basically the same quality than the input image (Fig. 2.33).

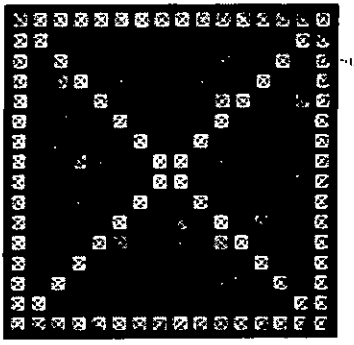


Fig. 2.33: Test pattern, reflected from the LCLV read side, observed at the output of LCTV2. The image quality of the fed back signal is as good es the quality of the signal coming directly from the input (cf Fig. 2.32.c).

2.7.2 Throughput

As we will see in section 3.5.3, the intensity required on the write side of the LCLV is ≥ 15 μW/cm². Write intensities up to 175 nW/cm² per activated input pixel were measured with the initial version of the setup (illuminated microlens array at the input, Ar^+-laser output = 900 mW). This results in only 8.75 μW/cm² for an input image with ~50 correct input pixels. In fact, we realized that only input images with ≥ 100 correct pixels could activate the LCLV (when operated at highest sensitivity).

To improve this situation we tried two things: reduction of the losses and combination of the LCLV with an image intensifier in order to rise the write light sensitivity. The latter was not successful, because of the limited brightness of the phosphor layer on the output side of the image intensifier and because of difficulties in efficiently imaging the intensified image from the diffuse (Lambertian) intensifier output to the LCLV write side. The conclusion of this test was that a combination of image intensifier and LCLV is only feasible if the glass substrate on the write side of the LCLV is replaced by a fiber optic plate that can be brought into contact with the intensifier output. Because we did not wish to open the LCLV, this has not been tested experimentally.

The reduction of the losses was done by replacing the illuminated microlens array by a Dammann-grating and a lens, as described in section 2.3. With this improvement, we were able to activate the LCLV even with only a few active input pixels (cf. Fig. 4.7.d, section 4.3.1). The intensity on the write side of the LCLV has been raised from 175 nW/cm² to about 1.35 μW/cm² per activated input pixel (Ar^+- laser output = 900 mW).

The size of one output pixel is 26 times the size of a LCTV2 pixel (cf. section 2.6.3), i.e. 26·21.5 μm x 26·31.5 μm = 0.004578 cm². An intensity of 1.35 μW/cm²

corresponds thus to a power of ~6.2 nW per output channel and per activated input pixel. At the input of the system, we measured ~675 µW per channel (or pixel). The signal has thus been attenuated by a factor of ~108'870. This value is composed of two contributions: a factor of 256 due to the 16 x 16 fan-out at DG2, and a factor of ~425 due to losses of the optical system.

Sufficient loop gain for a stable system state and even self-activation has been demonstrated (cf. section 4.3.2).

2.8 Optimization potential

Further optimization and miniaturization of the system is certainly possible: VCSEL arrays can replace the large input part (DG1 and L1). Specialized optomechanics (slotted baseplate or a custom-designed monoblock) instead of off-the-shelf mounts will allow further downscaling. The availability of higher resolution LCTVs (or other suitable spatial light modulators) will allow a higher number of channels.

The slow and non-uniform LCLV (cf. chapter 3) can be replaced by a suitable smart pixel element, such as a sandwiched combination of a detector array, electronics and a VCSEL array. This would considerably accelerate the cycle time of the feedback loop. Enhanced functionality of the neuron-plane and of course better uniformity are additional advantages of such a setup.

A different operation wavelength or a downsizing of the system will require a redesign of the setup and as a consequence a complete new set of hardware. A much easier improvement can be achieved in the short term by replacing the LCLV with a custom made smart-pixel device that has the 1092 µm channel pitch of this system.

Integration of VCSEL technology and miniaturization is also interesting for the matrix-vector-multiplier alone, which is an universal processing module that can be useful for applications other than optical neural networks. The interlaced fan-out method that has been introduced in section 2.5.2 offers an elegant way to realize a large number of parallel channels with minimal imaging errors. The use of a dilute VCSEL array at the input of such a module is interesting in the context of the heat related problems of VCSELs.

3 Optical neurons - The liquid crystal light valve (LCLV)

3.1 Overview and Introduction

In this chapter, we will have a closer look at the liquid crystal light valve (LCLV). In a system with all-optical feedback, the LCLV acts as an array of neurons and is therefore a key device that has to be understood in detail.

A main criterion for the evaluation of the device was the achievable gain. Our choice was finally made for a LCLV from the P.N. Lebedev Physics Institute in Moscow, Russia. This element is in principle well suited for our setup, but after a while, we realized a severe limitation: the insufficient homogeneity of the liquid crystal layer. This has been a known problem with liquid crystal devices for a long time [9, 10] and it remained one throughout this project.

After the Lebedev Institute stopped its LCLV-activity, we looked around for alternative devices, but throughout the period of experimental work, we were not able to find a suitable replacement. Having no alternative, we nevertheless characterized the device for its use in our recurrent neural network. Some of the following results are of a more qualitative nature, because the insufficient uniformity introduced considerable variations to some of the absolute measurements. Not everything was examined to the last detail because of the lack of general validity of the results.

In this chapter, the structure and operation principle of the LCLV will be introduced in sections 3.2 and 3.3. Experimental results of the characterization of the device are presented in sections 3.4 to 3.8.

3.2 Structure of the LCLV

The LCLV consists of several layers as shown in Fig. 3.1.a [39]. A photo of the device is shown in Fig. 3.1.b.

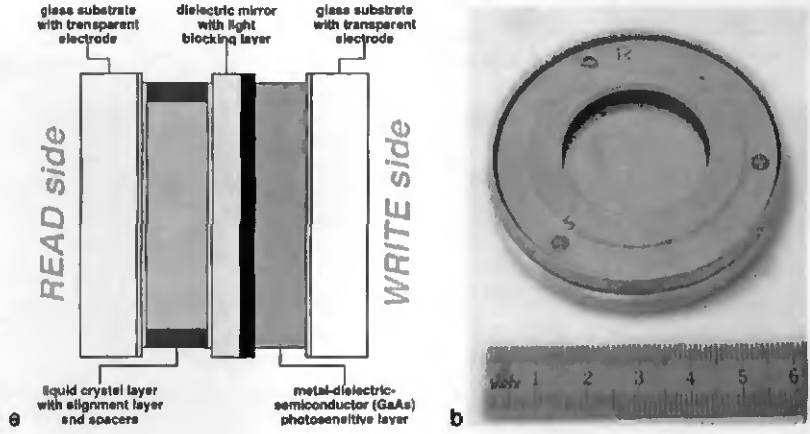


Fig. 3.1: a) Schematic sketch of the liquid crystal light valve structure. b) Photo of the device.

The structure is sandwiched between two glass substrates. Two transparent electrodes (standard indium-tin-oxide coating) allow to apply an electric field across the device. The clear aperture of the device is circular and has a diameter of 25 mm.

The liquid crystal layer is ~5 µm thick. A mixture with positive dielectric anisotropy near 8 and optical anisotropy of 0.14 is used as liquid crystal. Two unidirectionally rubbed polymer coatings are responsible for a parallel alignment of the nematic liquid crystal molecules. The pre-tilt angle is less than 2°.

The structure has a dielectric mirror, centered at 500-550 nm, and a light blocking layer that prevents the readout beam to interact with the write side of the device.

The photosensitive layer consists of a metal-dielectric-semiconductor structure with high-resistivity monocrystalline GaAs as the semiconductor.

3.3 Principle of operation

In short, the LCLV is a controllable wave retarder for a beam incident on its read side. The wave retardation can be varied globally by changing the applied drive signal or locally by changing the intensity distribution of a beam incident on the write side of the element. In both cases the electric field distribution across the liquid crystal is changed. This changes the tilt of the liquid crystal molecules which results

in a different wave retardation. When the device is used in combination with suitable polarizers, this change in wave retardation results in a change in readout intensity.

In the following, the principle of operation is discussed in more detail and with regard to our specific device. Further information about the optical properties of liquid crystals and liquid crystal devices can be found e.g. in [16, 40]. A general introduction to the physics of liquid crystals can be found e.g. in [41].

3.3.1 Behavior of liquid crystal in an electric field (no write signal)

Without an external electric field applied, the molecules of the liquid crystal align in the direction of the alignment layer (i.e. parallel to the LCLV surface) due to intermolecular forces.

When an electric field is applied, electric dipoles are induced because of the elongated shape of the liquid crystal molecules. The resultant electric forces exert a torque on the molecules and the molecules rotate in a direction such that the sum of elactrostatic energy (due to the electric field) and elastic energy (due to intermolecular forces) is minimized.

In our case, an electric field in the z-direction is generated by applying an AC signal with amplitude V across the two transparent electrodes. The resultant electric forces tend to tilt the molecules toward alignment in the z-direction, but the elastic forces at the surfaces of the alignment layers resist this motion. When the applied field is sufficiently large ($V > V_c$), most of the molecules tilt, except those adjacent to the glass surfaces. The equilibrium tilt angle Θ for most molecules is a function of V, which can be described by [41]

$$\Theta = \begin{cases} 0, & \text{for } V \le V_c \\ \frac{\pi}{2} - 2\tan^{-1}\exp\left(-\frac{V - V_c}{V_0}\right), & \text{for } V > V_c \end{cases}, \qquad (3\text{-}1)$$

where V is the applied rms-voltage, V_c a critical voltage at which the tilting process begins and V_0 a constant (cf. Fig. 3.2). When the electric field is removed, the orientations of the molecules near the glass surfaces are reasserted and all of the molecules tilt back to their original orientation.

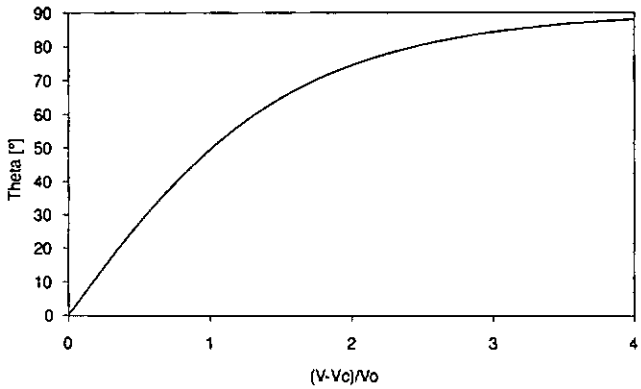


Fig. 3.2: Tilt angle of the liquid crystal molecules in function of the normalized voltage across the liquid crystal cell (according to Eq. 3-1).

Apart from the *dielectric* anisotropy, which is responsible for the tilting process, liquid crystals also show *optical* anisotropy. In our case of a nematic liquid crystal, the cell behaves like an uniaxial crystal. The normal modes of an optical wave travelling in the z-direction are polarized parallel and perpendicular to the orientation of the liquid crystal molecules. The associated refractive indices are the ordinary index n_o and the extraordinary index n_e. A cell of thickness d provides a wave retardation

$$\Gamma = \frac{2\pi}{\lambda}(n_e - n_o)d \quad . \tag{3-2}$$

When an external electric field is applied, the retardation becomes

$$\Gamma(\Theta) = \frac{2\pi}{\lambda}(n(\Theta) - n_o)d \quad , \tag{3-3}$$

where $n(\Theta)$ is given by

$$n(\Theta) = \left(\sqrt{\frac{\cos^2\Theta}{n_e^2} + \frac{\sin^2\Theta}{n_o^2}} \right)^{-1} \quad . \tag{3-4}$$

The retardation $\Gamma(\Theta)$ is maximal when $\Theta = 0°$ (no field applied, molecules not tilted) and decreases towards 0 when the tilt angle reaches 90°, i.e. when all the molecules are oriented in the z-direction (Fig. 3.3).

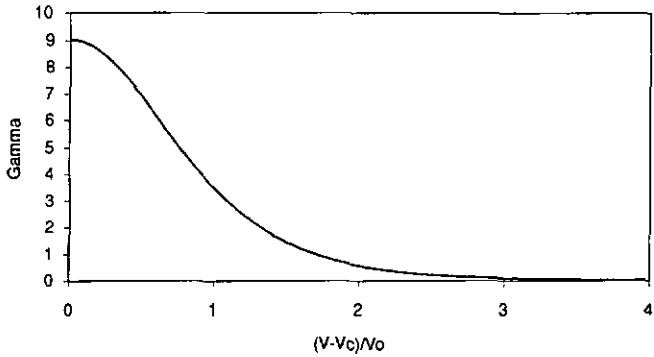


Fig. 3.3: Wave retardation provided by one passage through the liquid crystal cell in function of the normalized voltage across the liquid crystal (following Eq. 3-3, 3-4 and 3-1).

In the case of our device, we have the following values (according to the data sheet that came with the device): thickness of the liquid crystal layer $d = 5$ µm, wavelength $\lambda = 488$ nm and optical anisotropy $n_e - n_o = 0.14$. With these parameters we get a maximal retardation of $\Gamma_{max} = 9.0 = 2.9 \cdot \pi$.

Apart from the *retardation* Γ, which is given by multiples of π, the term *retardance* ρ is often used, which expresses the retardation in terms of the wavelength. The relation between the two terms is given by

$$\Gamma = \frac{2\pi}{\lambda} \rho \quad , \tag{3-5}$$

i.e.

$$\rho = (n(\Theta) - n_o)d \quad . \tag{3-6}$$

For our device, the maximal retardance $\rho_{max} = 1.43 \cdot \lambda$ (no voltage applied).

The liquid crystal cell alone can be used as a voltage controlled wave-retarder. In combination with suitable polarizers, such a wave retarder can act as an intensity modulator. Our device incorporates a mirror and is always used in combination with a polarizing beamsplitter (PBS). Since we are mainly interested in the intensity modulation properties, we shall consider the polarizing beamsplitter as an integral part of the LCLV. The term "LCLV output" refers thus to the intensity distribution observed at the deviated output of the polarizing beamsplitter (Fig. 3.4).

Up to now, the retardance has been discussed for one single passage through the liquid crystal cell. Because of the mirror in the LCLV, we finally have to deal with a double passage through the liquid crystal cell. In the following discussion, the terms

"cell retardance" and "LCLV retardance" are used for the retardance after a single passage through the liquid crystal cell and for the accumulated retardance after two passages through the liquid crystal cell, respectively.

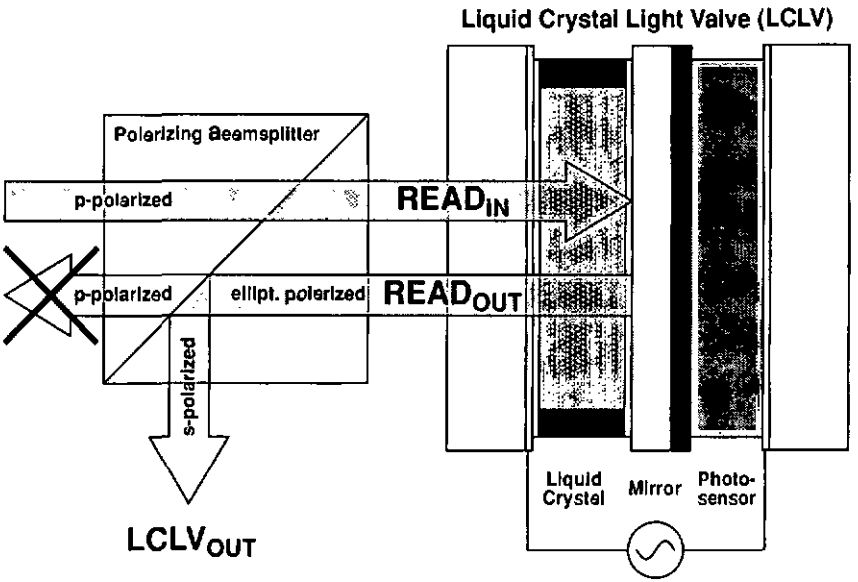


Fig. 3.4: Intensity modulation with the LCLV is achieved by using a polarizing beamsplitter and observing only the s-component of the reflected read$_{OUT}$ beam. The polarization of the read beam is modulated in function of the retardation provided by the liquid crystal. The retardation can be varied globally by changing the drive voltage or locally by a write beam incident on the photosensor.

In order to get maximal birefringence, the angle between the polarization direction of the polarizing beamsplitter and the rubbing direction of the alignment layer is set to 45°. The light passing straight through the beamsplitter is p–polarized (i.e. horizontally polarized; parallel to the breadboard surface). In the general case of an arbitrary retardance, the polarization of the reflected read$_{OUT}$ beam will be elliptic. Only the s-component (vertically polarized; perpendicular to the breadboard surface) is observed at the LCLV output (cf. Fig. 3.4). In the special case of a cell retardance of $\lambda/4$, we get circular polarization incident on the dielectric mirror. Reflection on the mirror changes the sense of the circular polarization, and after the inverse passage through the liquid crystal cell, the accumulated LCLV retardance is $\lambda/2$. This results in a rotation of the incident polarization of $\pi/2$, i.e. the light reflected from the LCLV is completely s–polarized. This corresponds to the maximum LCLV output (or *high* output). A cell retardance of $\lambda/2$, on the other hand, corresponds to a LCLV retardance of λ, and results in a completely p-polarized read$_{OUT}$ beam. This corresponds to zero LCLV output (or *low* output).

A more general and quantitative description of the relations between retardance, polarization and resulting intensity is obtained by using Jones-matrix calculus [40, Appendix C].

3.3.2 Influence of write signal

So far, only *global* variations of the applied electric field have been discussed. By applying a light intensity distribution on the write side of the valve, chargé carriers are produced in the semiconductor, which result in a spatial charge distribution and thus a *spatial* modulation of the electric field. For a constant drive signal amplitude, a local decrease of the impedance on the photosensor side results in a local increase of the voltage across the liquid crystal. According to Eq. 3-1, the molecules of the liquid crystal start to tilt locally, if the respective voltage change is large enough. It is therefore possible to modulate an uniform $read_{IN}$ signal with a suitable write signal. Fig. 3.5 shows a schematic sketch.

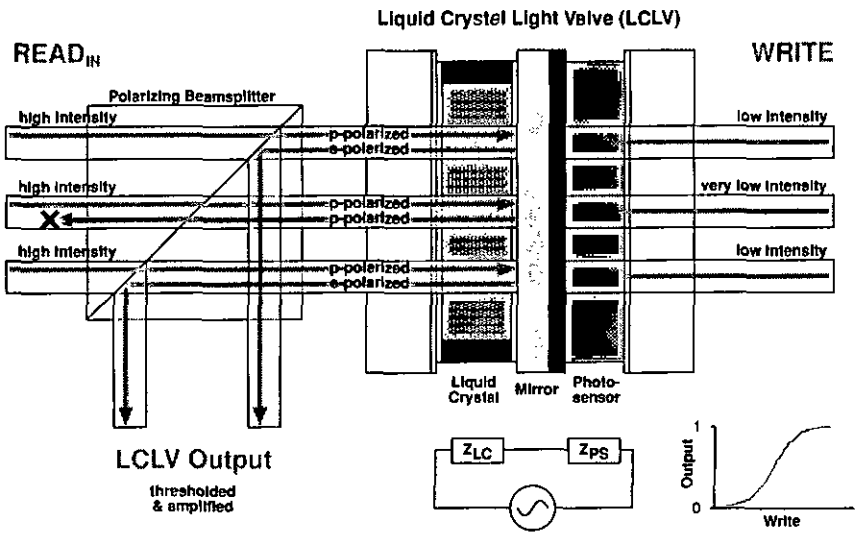


Fig. 3.5: A light intensity distribution on the write side of the LCLV results in a local variation of the voltage across the liquid crystal and therefore a locally different tilt of the liquid crystal molecules. The write pattern can therefore be modulated on the read beam. The nonlinear transfer-function of the LCLV results in a discrimination of the weak write signals (thresholding property of the LCLV) and the higher intensity of the read beam allows for an amplification of the signal.

The relation between write light intensity and read-out intensity is nonlinear (cf. section 3.6), which makes the LCLV a suitable *thresholding device* for optical neural networks. In our case of a *recurrent* neural network the LCLV also serves to *refresh* the signal which circulates in the feedback loop. In function of the quality of the light blocking layer, the read beam intensity can be several orders of magnitude higher than that of the write beam. The resulting gain is necessary to compensate for absorption and losses in the feedback loop.

3.4 Dark-state operation mode

To use the full dynamic range of the LCLV in our application, we want to drive it in such a way that the LCLV output is zero when no write beam is applied. The LCLV output intensity should increase as soon as the photosensor is exposed to a write beam of weak intensity. This operating mode is called *dark-state-mode*. It corresponds to a retardance of the liquid crystal cell of $n\lambda/2$ ($n \in N_0$) and thus a p–polarized $read_{OUT}$ beam (cf. section 3.3). The only free parameters are the amplitude and the frequency of the AC drive signal. Fig. 3.6 shows a map of possible amplitude-frequency-pairs that result in a dark-state operating mode.

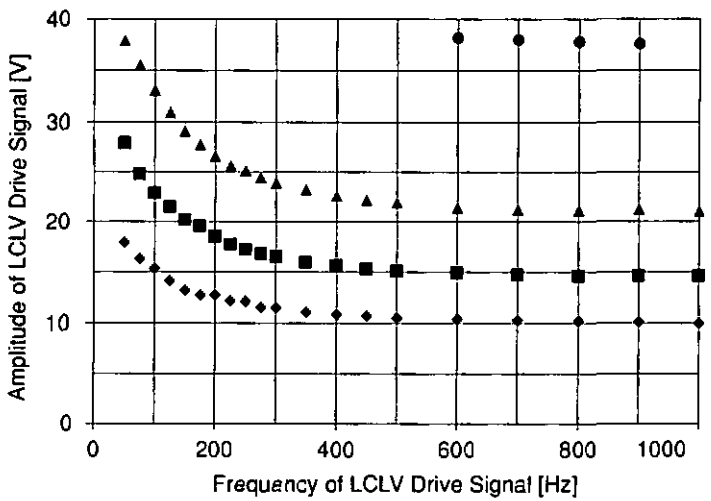


Fig. 3.6: Possible amplitude-frequency pairs of the LCLV drive signal that result in a dark-state operation mode. Each curve corresponds to a wave retardance of the liquid crystal cell of a multiple of λ/2, resulting in a p-polarized reflected beam.

Fig. 3.7 shows a plot of the normalized LCLV output intensity for various LCLV drive signal amplitudes between 0 and 40 V, at one fixed frequency of 5 kHz. The normalized LCLV output never reaches zero because of the uniformity problems that will be described in section 3.8 (this measurement has been made using the whole active surface of the device). The minima of this curve are possible dark-state modes.

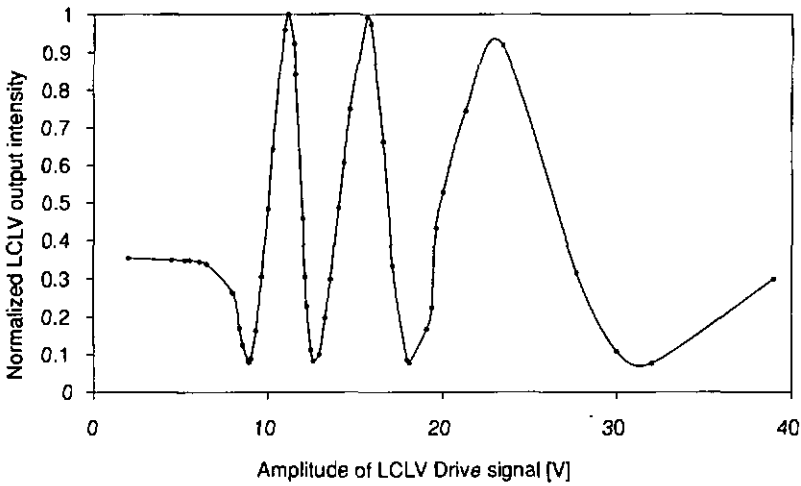


Fig. 3.7: Normalized LCLV output intensity in function of the amplitude of the LCLV drive signal, at a fixed drive frequency of 5 kHz. A 4th maximum has been observed at ~80V. See also Table 3.1.

We assume that a complete tilt of the liquid crystal molecules is achieved for a drive signal amplitude ≥ 100 V. The maximum at ~80 V corresponds to a LCLV retardance of $\lambda/2$, the minimum at 32 V to a LCLV retardance of λ, and so on. The maximum LCLV retardance is thus $\sim 4.2 \cdot \lambda$, which corresponds to a cell retardance of $\sim 2.1 \cdot \lambda$. From the values provided on the LCLV data sheet, however, we expect a cell retardance of $1.43 \cdot \lambda$ (cf. above). Assuming a thickness $d = 7.3\ \mu m$ instead of 5 µm in Eq. 3-6 would explain the discrepancy. See also section 3.8 for further remarks about the LCLV cell thickness.

Table 3.1 gives a summary of various parameters for the maxima and minima of Fig. 3.7.

drive voltage	cell retardation	cell retardance	LCLV retardation	LCLV retardance	res. output polarization	LCLV out (PBS)
0 V	$4.2\cdot\pi$	2.1 λ	$8.4\cdot\pi$	4.2 λ	ellipt.	0.25
9 V	$4\cdot\pi$	2λ	$8\cdot\pi$	4 λ	p	0
11 V	$3.5\cdot\pi$	7λ/4	$7\cdot\pi$	3.5 λ	s	1
13 V	$3\cdot\pi$	3λ/2	$6\cdot\pi$	3 λ	p	0
16 V	$2.5\cdot\pi$	5λ/4	$5\cdot\pi$	2.5 λ	s	1
18 V	$2\cdot\pi$	λ	$4\cdot\pi$	2 λ	p	0
24 V	$3\cdot\pi/2$	3λ/4	$3\cdot\pi$	1.5 λ	s	1
32 V	π	λ/2	$2\cdot\pi$	λ	p	0
(80 V)	π/2	λ/4	π	λ/2	s	1
(>80 V)	0	0	0	0	p	0

Table 3.1: Various LCLV parameters for the maxima and minima of Fig. 3.7.

3.5 Achievable gain

As mentioned above, the achievable gain was one of the main criteria for the choice of the LCLV. The gain G of the activated valve can be defined as the ratio between the LCLV output intensity and the minimal intensity of the write beam, which is needed to activate the valve, i.e.

$$G = \frac{I_{LCLV_{OUT}}}{I_{write_{min}}} = \frac{\eta \cdot I_{read_{IN}}}{I_{write_{min}}} \quad , \tag{3-7}$$

where η is a constant factor, taking into account the reflectivity of the dielectric mirror, the absorption of the liquid crystal layer and the efficiency of the polarizing beamspliffer (cf. section 3.5.1). From this formula it can be seen that the gain is proportional to the $read_{IN}$ beam intensity, which is in practice limited by a) the available laser power and b) by the quality of the light blocking layer between read and write side of the LCLV (cf. section 3.5.2). If this layer is not efficiently blocking, the $read_{IN}$ beam starts to penetrate to the write side at a sufficiently high intensity, and the LCLV will be self-activated, regardless of the write signal. A bad light blocking layer results therefore in a limitation of the possible $read_{IN}$ beam intensity and thus a limitation in achievable gain.

3.5.1 Reflectivity and absorption

The dielectric mirror of our LCLV is centered at 500 - 550 nm, according to the data sheet of the device. The available Ar^+-laser, however, had its maximal emission at 488 nm. Scanning a range of wavelengths between 400 and 700 nm by means of a monochromator, the intensity of the reflected beam was compared to the intensity of the incident beam (Fig. 3.8). Unpolarized light was used for these experiments.

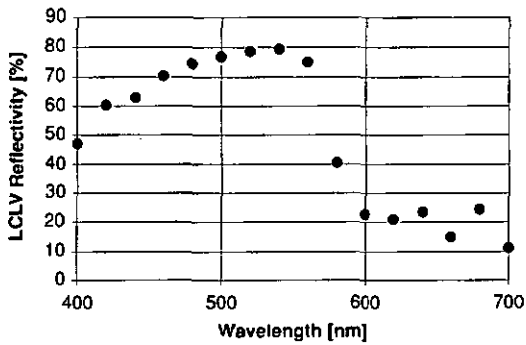


Fig. 3.8: Reflectivity of the LCLV at various wavelengths. These values include losses (scattering and absorption) caused by the liquid crystal cell in front of the mirror.

The absolute values in Fig. 3.8 include the losses (scattering, absorption) caused by the liquid crystal layer just in front of the mirror. The reflectivity of the LCLV at 488 nm is 75%, which is acceptable for our application.

The double passage through the polarizing beamsplitter causes additional losses of about 7 %. This results in an efficiency factor $\eta \approx 0.7$ for Eq. (3-7).

3.5.2 Quality of the light blocking layer

To test the quality of the light blocking layer, we operated the device in a dark-state mode without write signal and illuminated the read side with a beam of high intensity. In case of a breakthrough of the $read_{IN}$ beam to the write side, we expect to see an increase in the LCLV output intensity due to the self-activation of the LCLV.

We could apply $read_{IN}$ beams with an intensity of up to $6 \cdot 10^7$ $\mu W/cm^2$ (beam of 230 mW power and 0.7 mm diameter) without any observable breakthrough to the write side. Even higher values might be possible, but this was not tested because of limited available output power of the laser.

3.5.3 Write light sensitivity

As mentioned above, the write light sensitivity is the second factor that determines the gain of the device. We define the write light sensitivity as follows: $I_{write-min}$ is the minimal intensity that is needed on the write side to switch the LCLV output from zero to 90 % when the valve is operated in a dark-state mode. Fig. 3.9 shows experimental data of $I_{write-min}$ vs. the drive signal frequency.

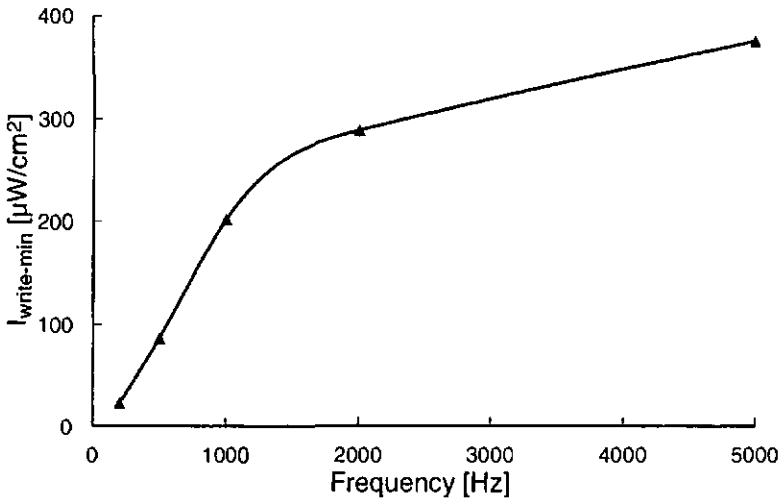


Fig. 3.9: Measurement of the minimal write intensity that is needed to activate the LCLV (0 to 90 %) for different drive signal frequencies, when the LCLV is operating in a dark-state mode.

For lower drive signal frequencies, less light is needed to activate the valve. For a constant $read_{IN}$ beam intensity, this allows to control the gain via the frequency of the drive signal. The lowest measured write intensity which can still activate the LCLV was 15 µW/cm² (at a drive signal frequency of 150 Hz).

The relation between drive frequency and LCLV sensitivity can be explained by mechanisms in the semiconductor of the photosensor, which is only photosensitive during a certain time of the drive signal cycle. A detailed description is given in [42]. The same mechanism is also responsible for the relation between drive signal frequency and spatial resolution of the LCLV.

3.5.4 Conclusion for the echieveble gein

Having determined all parameters, one can now calculate the achievable device gain using Eq. (3-7). With $\eta = 0.7$, $I_{read\text{-}IN} \geq 6 \cdot 10^7$ µW/cm² and $I_{write\text{-}min} = 15$ µW/cm², we obtain $G \geq 2.8 \cdot 10^6$.

The device gain represents an upper limit. More important in practice is the effective gain, which is smaller than the device gain for a number of reasons: a) The available beam power is reduced by spatial filtering, reflections, straylight and absorptions in polarizers, etc. b) Only about half of the total laser power is available for the read beam when the same laser is used for illumination of the write side. c) The LCLV is used as an array of individual neurons. Distribution of the available beam power into the total number of N individual channels (neurons) reduces the available $read_{IN}$ intensity per channel.

In our setup, we measured an available $read_{IN}$ beam power $P_{read\text{-}IN} \leq 675$ µW per channel. The minimal intensity incident on the LCLV write side which is required to activate the LCLV is $I_{write\text{-}min} = 15$ µW/cm². With a pixel (channel) area at the LCLV write side of about 0.004578 cm² (26 times the LCTV2 pixel size), this intensity corresponds to a power of ~68.7 nW per channel. Using Eq. 3-7, these values yield an effective gain of ≤ 7000.

In order to estimate if stable system states without external stimulus are possible, one has to calculate the loop gain. To do this, one has to take into account that the signal of one channel is fanned-out 16 x 16 times at DG2, i.e. the effective gain has to be divided by 256. A loop from the neuron output (LCLV read side) to the neuron input (LCLV write side) has therefore a gain of ~27. The absorption of such a loop is ~425 (cf. section 2.7.2). From this, we conclude that at least 16 simultaneously active channels are necessary for stable system states without external stimulus. See also section 4.3.2 for experimental results.

3.6 Transfer characteristics

Once a suitable combination of drive voltage and frequency has been fixed, the LCLV retardance will be determined locally by the intensity of an incident write signal. A sigmoid-shaped transfer characteristic can be observed (Fig. 3.10).

Fig. 3.10 is of course closely related to Fig. 3.7. In the first case (Fig. 3.7), a global voltage change results in a global change of the LCLV retardance. In the second case (Fig. 3.10), the change is only local and induced by an incident write beam. Fig. 3.10 actually represents a section of Fig. 3.7 from one minimum to the next maximum. Thus, it is obvious that an additional local LCLV retardance exceeding $\lambda/2$ will reduce the readout intensity again. A very intense write signal can therefore

result in a low $read_{OUT}$ signal. The choice of the work point (drive signal amplitude and frequency) has therefore to respect the intensity range of possible write signals.

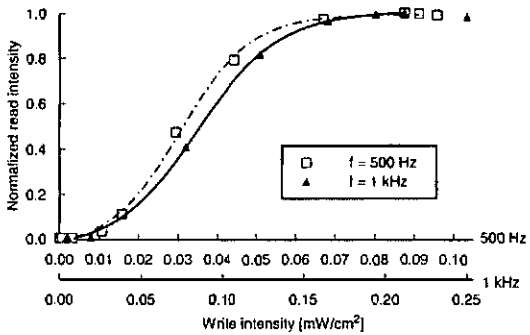


Fig. 3.10: Transfer charecteristic of the LCLV for two different drive frequencies.

From the above, it is also clear that it is possible to drive the LCLV in an inverse mode, i.e. the $read_{OUT}$ intensity is maximal when no write beam is applied. With increasing write intensity, the $read_{OUT}$ intensity decreases. A demonstration of this operation mode is given in section 4.3.1.

3.7 Response speed

We tested the response of the LCLV to a step function incident on the write side. This was simply done by opening an optical shutter on the write side and measuring the resulting signal on the read side. We observed that the rise time increased with decreasing write intensity. This can be understood as a consequence of the dynamical behavior of a liquid crystal in a changing electric field. More generally expressed, the rise time increases when the voltage across the liquid crystal is close to the critical voltage V_c. Measured values for the rise time (10 % to 90 %) were in the range of 900 ms ($I = I_{sat}$) up to 7.8 s ($I = 0.34\ I_{sat}$). The drive frequency for these measurements was 500 Hz.

The decay time (90 % to 10 %) remained constant at about 200 ms. This is because the reorientation of the liquid crystal molecules with respect to the alignment layer depends only on the elastic forces and not on the applied field (see e.g. [43, p. 77]).

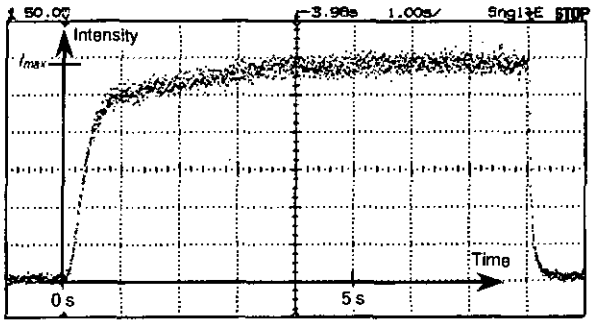


Fig. 3.11: LCLV $read_{OUT}$ response to a step function incident on the write side.

The dependence of the response speed on the write intensity is the reason for the asynchronous processing properties of the LCLV. This property certainly affects the dynamical behavior of the neural network. In reality the behavior is even more complex, because we do not have a simple step function at the input, but rather a continuous update of the write information, due to the very short propagation time (in the order of ns) of the signals from the read to the write side of the LCLV.

3.8 Uniformity of the liquid crystal layer

In spite of all the good properties, the LCLV had one severe limitation: an insufficient uniformity of its liquid crystal layer. This problem was not realized immediately. Most of the characterizations were made with beams of rather small diameter, and therefore only a small part of the device was illuminated at a time. After realizing significant variations between different sets of measurements, we had a look at the uniformity of the device. Fig. 3.12 shows a photograph of the LCLV readout side (uniformly illuminated and observed through a polarizing beamsplitter). The external voltage is zero and no write signal is applied. For a correctly working device, we would expect to see a uniform output signal. Our device, however, shows fringes. It its impossible to obtain a state of uniform behavior.

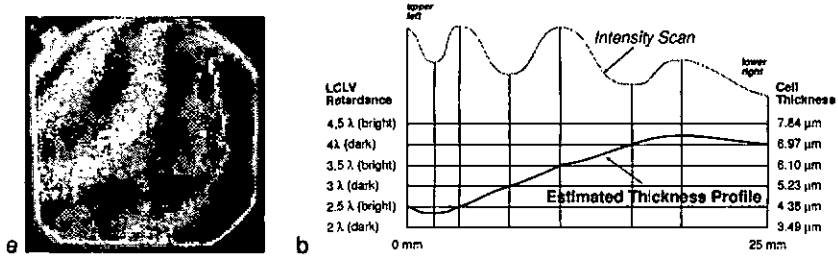


Fig. 3.12: a) The non-uniformity of the liquid crystal layer results in a fringe pattern at the LCLV output when the read side is illuminated with an expanded beam of uniform intensity. b) Intensity scan from the upper left to the lower right of Fig. 3.12.a and estimated thickness profile.

A thickness variation of the cell would explain the existence of the fringes. Looking at Eq. 3-2, one sees that the retardation of the cell also depends on its thickness. The fringes can be understood as regions of different retardance caused by a locally different cell thickness. Fig. 3.12.b shows a possible interpretation of the observed fringe pattern. The intensity scan and the number of fringes indicate a thickness variation of around 3 µm. The deviation between nominal thickness (5 µm, according to the data sheet) and experimentally determined thickness (~7.3 µm, indicated by the number of maxima and minima in Fig. 3.7), as described in section 3.4, is consistent with this interpretation.

As a result of this non-uniformity, it has been impossible to achieve a good dark-state-mode. The consequences for the use of the LCLV in the optical neural network will be described in section 4.2.

4 Test of the optical neural network

A chain is only as strong as its weakest element. In our case, the weakest element was clearly the liquid crystal light valve. As described in the previous chapter, we were not able to obtain an improved element.

As a consequence, most of the experimental results for the optical neural network are only of qualitative nature. It did not make sense to study the behavior of the system more quantitatively because the results were strongly influenced by the LCLV non-uniformity.

When looking at the images in the following sections, one should keep in mind that all the signals are superposed by LCLV noise. Otherwise, the bad quality of these images might give a wrong impression about the quality of the optical setup.

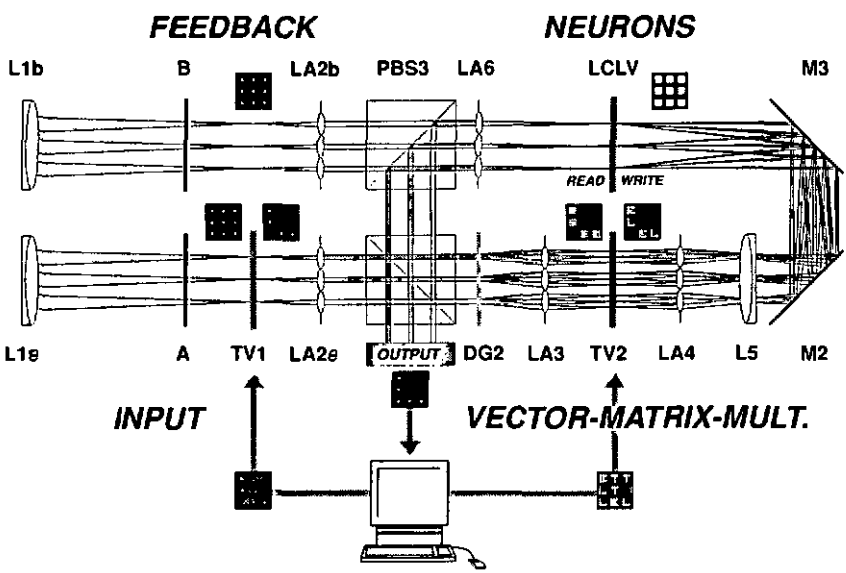


Fig. 4.1: Scheme of complete optical neural network.

We will begin this chapter with a short introduction to the simplified method which was used to calculate the weight matrix in the experimental demonstrations. The

influence of the *LCLV non-uniformity* to the system performance will be discussed in section 4.2, and the influence of the *LCLV work point* on the system behavior in section 4.3

4.1 Simplified weight matrix calculation

During the tests of the building blocks of our optical system (chapter 2), we developed a simplified method to calculate the weight matrix. This method was chosen because of its intuitive approach and because it allowed immediate control of the image formation in the optical system. That was especially useful for the correct matching of the feedback signals to the matrix-vector-multiplier.

The method is straightforward: each sub-array of the weight matrix is attributed to one input channel (interlaced fan-out). We then check every single pixel of a reference pattern (pattern that shall be stored in the weight-matrix) whether it is active or not. If yes, a copy of the pattern is placed in the sub-array corresponding to the pixel. The idea behind this method is obvious if one keeps in mind that the output of the matrix-vector-multiplier is a superposition of all the sub-arrays. A test pattern at the input that has 50 % of its pixels covered will still activate 50 % of its representations in the weight-matrix. Evidently, other patterns that share common pixels with the test pattern will be activated as well. Their activation, however, will be much smaller if the reference patterns are not strongly correlated. Illustrations of such kind of weight matrices can be found e.g. in Fig. 2.17, Fig. 2.20 and Fig. 2.24. Fig. 4.2 and Fig. 4.6.b. show the weight matrix used in the experiments that are presented in this chapter.

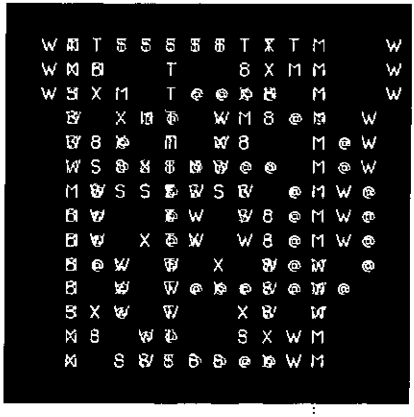


Fig. 4.2: Example of a weight matrix with encoded letters @, 8, M, S, T, W and X.

Our method works best for single pass recognition of partially covered input (associative recognition). No mathematically solid comparison (speed of convergence, system stability, storage capacity, etc.) with the original Hopfield algorithms has been made, but we can be quite certain that the latter is much better optimized for convergence in applications with large data sets and noise.

For convenience, and because a thorough study of the neural network behavior was hindered by the LCLV non-uniformity, we continued to use this model for the rest of the experimental work. As we can see in the following sections, it performed amazingly well, in spite of its simplicity.

4.2 System noise - Influence of LCLV inhomogeneity

Because of our system design, which minimizes imaging errors by keeping the channel pitch constant throughout the whole system, the size of the neural array was ~17.5 mm x 17.5 mm. We therefore have to use the whole active area of the LCLV. The insufficient uniformity of the liquid crystal layer across this large area makes it impossible to obtain a good (uniform) dark state mode (Fig. 4.3.a), which results in channels that are always activated (Fig. 4.3.b). A clean signal incident on the write side of the LCLV (Fig. 4.3.c, matrix-vector-multiplier output) will therefore appear noisy at the output of the LCLV (Fig. 4.3.d).

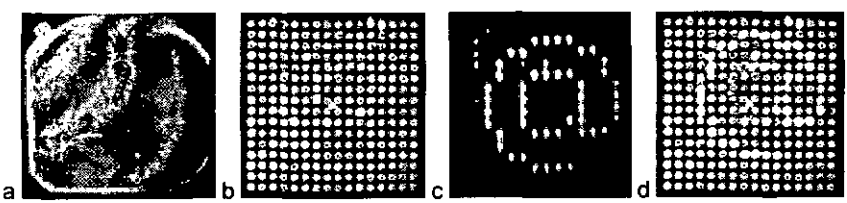


Fig. 4.3: Influence of the LCLV non-uniformity: a) A good, uniform dark-state operation mode is not possible. b) When the LCLV is used in the optical neural network, some channels are always active due to the non-uniformity. c-d) This causes the clean output of the matrix-vector-multiplier (c, incident on the write side of the LCLV) to appear noisy at the read side of the device (d, at the output of the polarizing beamsplitter). (Images of early test phase)

In addition, the channels that are always active cause a constant incorrect illumination ("idle noise") of some weight matrix elements, which results in even more incorrectly activated channels. Fig. 4.4 illustrates the influence of this idle noise: Fig. 4.4.a is the output of the matrix-vector-multiplier for a partially covered @-sign at the input (cf. Fig. 4.6.a). Fig. 4.4.b shows the matrix-vector-multiplier output when the

shutter A (cf. Fig. 4.1) is closed (no input) and shutter B (feedback beam) is open. The LCLV write side has been covered, i.e. this image should be black for an ideal dark-state. The non-ideal dark-state and straylight from the liquid crystal layer, however, result in a constant illumination level which incorrectly activates several patterns. Fig. 4.4.c shows the matrix-vector-output when both shutters are open: The clean signal of the input completely disappears in the noise.

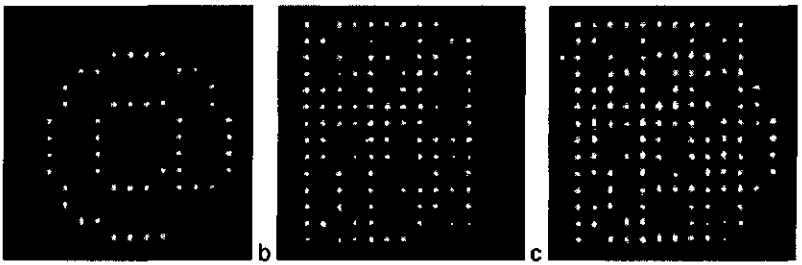


Fig. 4.4: Demonstration of the effect of idle noise, observed at the output of the matrix-vector-multiplier. The LCLV write side has been covered to isolate the effect of idle-noise. a) signal produced by a partially covered @-sign at the input (only shutter A open); b) idle-noise caused by the LCLV non-uniformity and straylight (liquid crystal layer) incorrectly activates several channels (only shutter B open); c) input signal together with noise from feedback (both shutters open).

If one looks more closely at the images in this section, one discovers another problem of the LCLV non-uniformity: the noise pattern is not the same in different figures. This has two reasons. 1) We do not use the whole LCLV surface but only 16 x 16 small sections distributed across the surface. A minor displacement of this array between measurements (removal of the LCLV for other measurements, improvements to the optical setup, etc.) will lead to a different noise pattern. 2) The dark state operation point cannot exactly be determined, because of the non-uniformity, and therefore the reproducibility is limited. Fig. 4.5 illustrates this effect: three slightly different operation points around the dark state operation point produce different noise patterns. Their overall darkness (criterion for a useful dark state) is about equal.

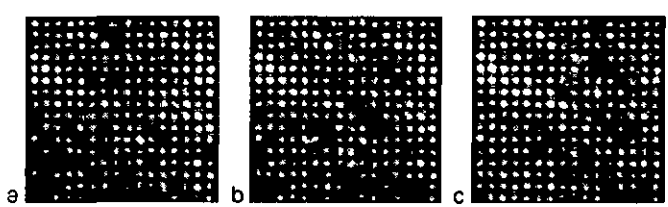


Fig. 4.5: LCLV output for different operation points around the dark state point. The difficulty in determining the dark state results in bad reproducibility.

Fig. 4.6 shows another example of a pattern recognition demonstration. Please note that the weak spots in Fig.;4.6.c are not due to noise; this image represents the unthresholded output of the matrix-vector-multiplier. The weak spots are thus due to other patterns stored in the weight matrix, which have pixels in common with the input signal. The additional spots in Fig. 4.6.d, however, are due to LCLV noise. These images represent the lowest noise level which has been achieved during experimental work. The system was not operated in the feedback-mode during this particular test, but in a single-pass mode.

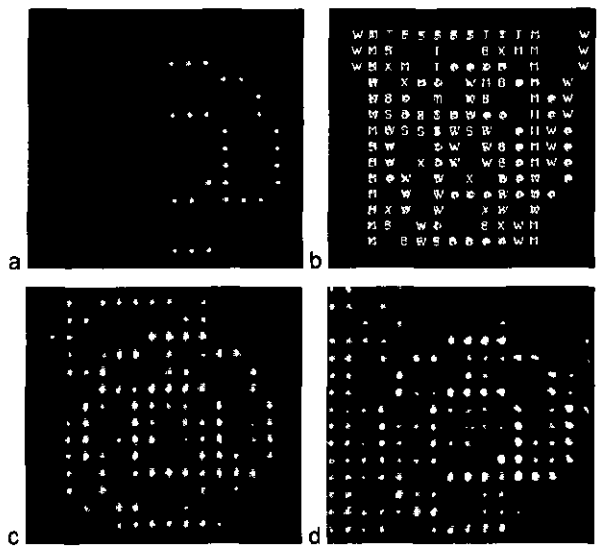


Fig. 4.6: Pattern recognition demonstration for a single-pass mode of the network. a) input pattern (LCTV1); b) weight matrix (LCTV2); c) unthresholded matrix-vector-multiplier output (LCLV write side); d) thresholded system output (LCLV read side). Additional spots on c) are due to other stored patterns, which share pixels with the input image; additional spots on d) are due to LCLV noise.

The (partial) compensation of system imperfections by adjusting the weight matrix during an in-situ training of the system is in principle possible thanks to the reconfigurable weight matrix and the control hard- and software. However, one has to keep two points in mind: 1) Any such compensation of systematic errors will consume a part of the available contrast of the LCTV, which in turn reduces the available contrast for pattern recognition. 2) The amount of noise introduced by our particular LCLV is far beyond of what can be called slight homogeneity variations. A detailed experimental study of the error compensation power of this system has not been made.

4.3 Influence of the LCLV work point on the system behavior

One of the main concerns for future work will be the working point of the LCLV. We realized that the values obtained during the element characterization are not easily transferable to the complete system. One reason is that the intensity on the write side of the LCLV (output of vector-matrix-multiplier) not only depends on the absolute beam intensities but also on the test patterns. Patterns with many active pixels result in a higher write intensity than patterns with only few pixels. An ideal working point will thus also depend on the patterns to be recognized. With our control hard- and software we have in principle the possibility to change the LCLV working point during the operation of the neural network. The successful implementation of such a feature, however, requires a careful analysis of the resulting system behavior.

After optimization of the input part (cf. section 2.7.2), we had enough write light intensity to experimentally test several working points for the LCLV (cf. section 3.4, Fig. 3.6). The choice of a working point determines on one hand the threshold level of the optical neurons and on the other hand it determines the gain of the feedback loop. In the following two sections, experimental demonstrations for both cases are presented.

4.3.1 Threshold level (without feedback)

Fig. 4.7 demonstrates the influence of the LCLV work point (and thus the LCLV loop gain) on the threshold level. A partially covered @-sign is applied at the input of the system (a). Fig. 4.7.b shows the output of the matrix-vector-multiplier (write side of the LCLV). One can already see the restored @-sign together with weaker pixels (coming from other patterns stored in the weight matrix that have common pixels with the input pattern). The task of the LCLV is now to discriminate between the weak and the strong pixels. If the gain is too low, the output signal of the LCLV will be zero (no picture); if the gain is too high, the LCLV output will contain all pixels, even the weak ones (c). Only if the LCLV gain is in between these two extremes, the desired pattern will appear at the output.

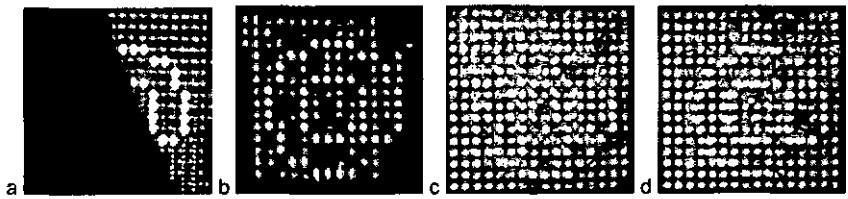


Fig. 4.7: Demonstration of the influence of the LCLV work point to the threshold level: a) Input signal (partially covered @-sign); b) Output of the matrix-vector-multiplier, incident on the LCLV write side; c) LCLV output with high gain: also the weak pixels of image B are activated; d) LCLV output with low gain: only the strong pixels of image B are activated. Additional pixels are due to LCLV noise.

Another interesting possibility can be seen in Fig. 4.8. Here, the LCLV is driven just in the opposite of a dark state configuration; when no write signal is present, all the output channels are on. When a signal is applied to the write side of the LCLV, the intensity of the corresponding output channels will decrease and we observe an inverted version of the output image.

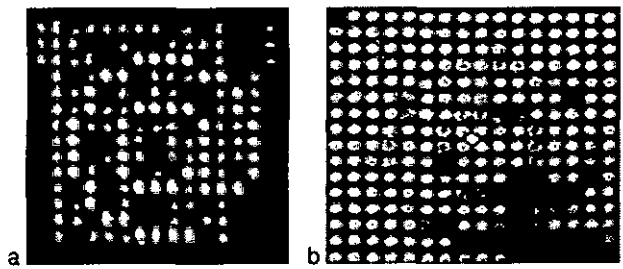


Fig. 4.8: It is also possible to drive the LCLV in an inverse mode, where all channels are ON if no write signal is applied. If a write signal (a) is applied, the intensity of the corresponding output channels decreases.

4.3.2 Loop gain and system stability (with feedback)

The LCLV working point will also determine the behavior of the network with regard to its memorizing capabilities. If the gain of the LCLV is set high enough, the system is capable to remain in a stable state after recognition of the test pattern, even when the input test pattern is not present anymore (shutter A closed, cf. Fig. 4.1). We observed that the stability of this state depends (among other parameters) on the

time interval during which the test pattern was presented. For some specific system parameters, it might also be possible that the system enters into a regime of oscillations between stored patterns (as it was observed by Xue [10]). However, no such effect has been observed so far for our system, which is probably a result of the massively increased number of neurons. The correlation between the 3 x 3 patterns of Xue was much stronger than the correlation between our 16 x 16 patterns. Moreover, the number of possible system states of our system is much higher (2^{256} compared to 2^9).

Fig. 4.9 to Fig. 4.11 show the influence of various values of the LCLV loop gain on the system stability. For very high gain, the system always enters into a saturated state, even without an input signal (Fig. 4.9). The idle-noise caused by the LCLV (Fig. 4.9.a) provides sufficient light to activate some of the patterns in the weight matrix. This in turn activates other channels and the system converges towards a saturated state (Fig. 4.9.f).

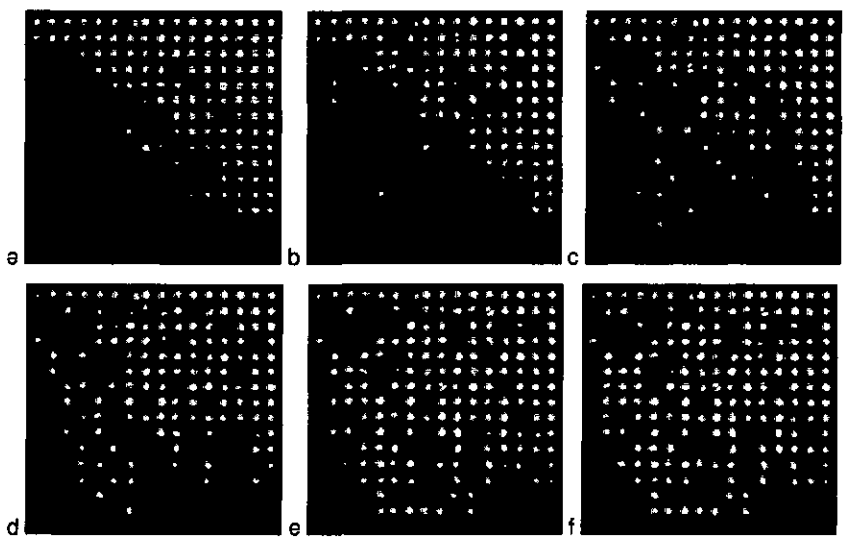


Fig. 4.9: Auto-saturation happens when the feedback loop gain is very high. The channels that are always active because of the LCLV noise (a) provide enough light to activate other channels (b-f). The steady-state pattern (f) depends on the noise pattern and on the patterns coded in the weight matrix.

For a lower, but still relatively high gain, there is no auto-saturation observed anymore. Fig. 4.10.a shows the idle noise. If a test pattern (partially covered @-sign, cf. Fig. 4.11.a) is applied at the input, the system reconstructs the stored reference pattern (Fig. 4.10.b-e). Thanks to the high gain, the state of the system is maintained, even if the input stimulus is removed (Fig. 4.10.f-i).

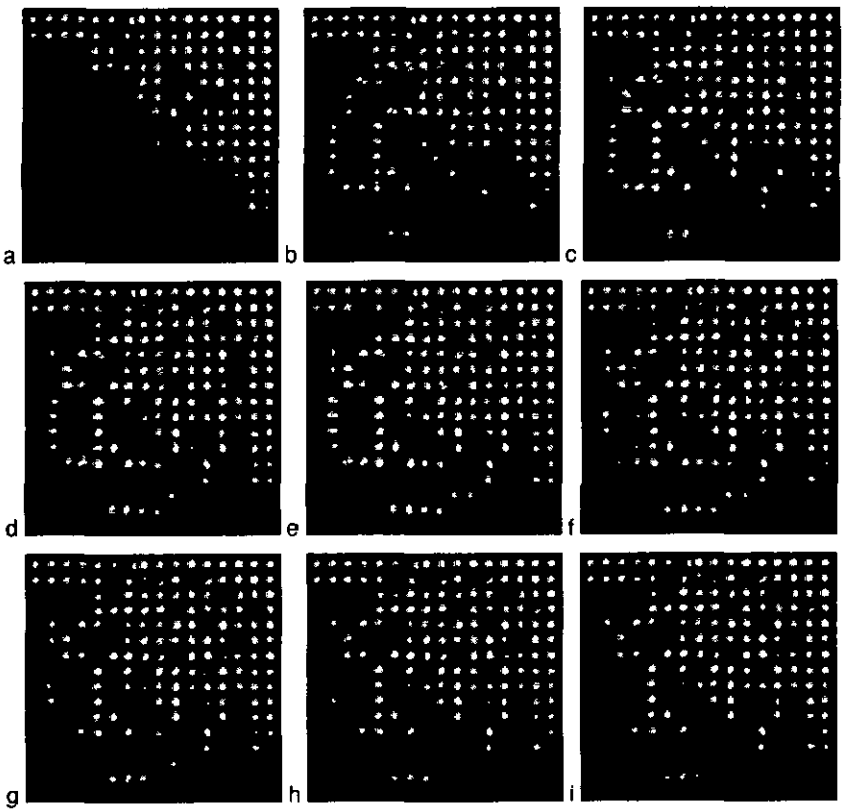


Fig. 4.10: With a moderate gain, there is no auto-saturation. a) Idle noise, caused by LCLV non-uniformity. b)-e) After presentation of a test pattern (partially covered @-sign, cf. Fig. 4.11.a), the system quickly converges to a state that corresponds to the reconstructed reference pattern. f)-i) The input stimulus is switched of. The loop gain is high enough to allow the system to remain in its state.

Fig. 4.11 shows the development for low gain. Fig. 4.11.a shows the test pattern presented at the system input (partially covered @-sign). Fig. 4.11.b shows the idle noise caused by the LCLV non-uniformity. After presentation of the input stimulus, the system converges to a stable state that corresponds to the reconstructed reference pattern (Fig. 4.11.c-e). When the external stimulus at the system input is switched off, the system returns to the idle state (Fig. 4.11.f).

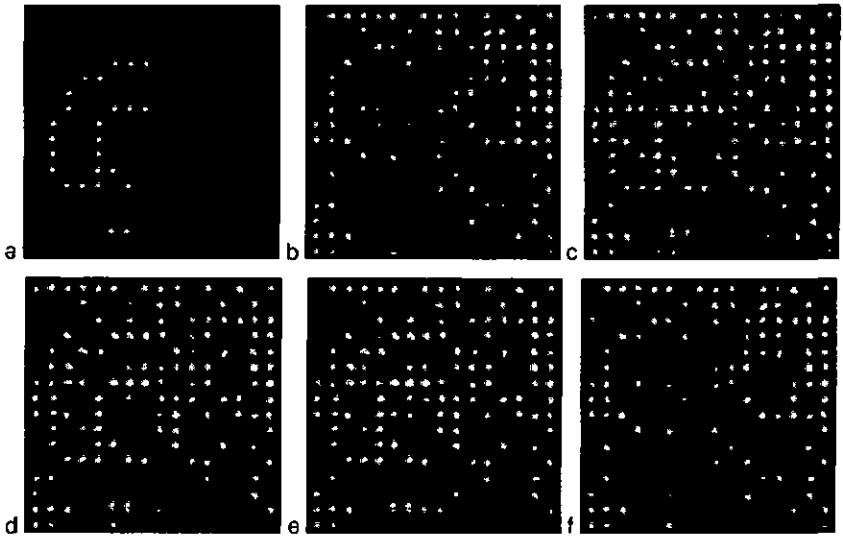


Fig. 4.11: Development of system state with low gain. a) Test pattern, presented at the system input (partially covered @-sign). b) Idle noise, caused by LCLV non-uniformity. c)-e) After presentation of the external stimulus (a), the system converges to a stable state that corresponds to the reconstructed reference pattern. f) When the external stimulus at the system input is switched off, the system returns to the idle state.

5 Summary and conclusions

We reported on the design, construction and testing of an optical neural network with 256 fully interconnected neurons, reconfigurable weight matrix and optical feedback loop.

The building blocks of the optical system have been discussed in the general context of massively parallel processing optical systems, with a focus on multistage systems with a large number of channels and thus large image fields. The considerations that lead to the final system design have been presented. The setup probes the physical and technological limits of some of its elements. Examples are the diffraction limited focusing and collimation with high f-number microlens arrays, low tolerances for image deformations over a large image field, tight requirements for the uniformity of microlens arrays and spatial light modulators over a large surface, and relatively low tolerance for losses, i.e. the need for high efficiency. The choice of the final system parameters was a compromise between numerous related and partially contradictory requirements.

The *optical design goals*, mentioned at the beginning of chapter 2, have all been reached. Compared to preceding all-optical networks, which have been realized at the Institute of Microtechnology [9, 10], the *number of channels* has been raised from 9 to 256, and the number of interconnections from 81 to 65'536. The fixed weight matrices have been replaced by *a reconfigurable weight matrix*. The final system size is larger than planned, but in spite of the increase in complexity and functionality, the system is still much smaller than its predecessors (by a factor of about 2-4). Further optimization of the setup is still possible. To our knowledge, the largest comparable all-optical neural networks on the international level are the ones of White [44] with 2 x 3 neurons, of Shariv and Friesem [45] with 4 x 4 neurons, and of Shariv *et al.* [46] with 8 x 8 neurons. All these networks have fixed weight matrices.

The improvements have been made possible by the use of novel methods and technology, namely the interlaced fan-out principle and the use of microlens-arrays. *The interlaced method of interconnecting each channel with all other channels* respects the complexity of different system stages in a better way than conventional methods. Imaging errors are minimized where the complexity is large, i.e. at the interconnection matrix (weight matrix). The microlens arrays allowed for the realization of microchannel-telescopes and (in combination with a macro-lens) for a true optical fan-in. Using microchannel-telescopes for the relay of discrete images, large image fields with low aberrations over the whole field are possible. A convincing demonstration of the potential of this approach are the 65'536 individual beams that all match their corresponding pixel apertures of the weight matrix LCTV. A constant channel pitch throughout the whole system is another advantage of the interlaced

method. It resulted in a robust setup which is easy to align in spite of the complexity of the system.

However, a chain is only as strong as its weakest element. In our case, the weakest element was clearly the liquid crystal light valve. The non-uniformity of its liquid crystal layer severely limits the performance of the complete neural network. There might be three possible solutions to overcome this problem:

1) Bad channels can be mapped out by means of the system control software. This of course reduces the number of neurons but it might allow for a more uniform neuron behavior.

2) A new LCLV which meets all the requirements might become available in the future (cf. below). The optical system is robust enough to allow the integration of a new device without a "realignment-nightmare".

3) Instead of using a LCLV with one extended liquid crystal cell, the use of a pixelated device with improved thickness uniformity should be considered. Such a device does not necessarily have to be a LCLV. It could for example be a sandwich consisting of a detector-array on the write side and a smart-pixel-array on the read side. The update of the cells on the read side can be synchronized or asynchronous.

The first of the three solutions is the most straightforward, because it can be realized immediately, provided the necessary manpower. The third method allows for additional system functionality, which might be useful for the test of novel neural network models (cf. below), but it requires the fabrication of a custom made device.

As a consequence of the LCLV non-uniformity, most of the experimental work with the completed system was only of qualitative nature. In particular, the influence of the LCLV work point on the system behavior has been demonstrated. With an improved array of neurons, it would be possible to explore the dynamical behavior of the system in more detail, to compare asynchronous and synchronous processing mode, to test several learning strategies (including in-situ training), and thus to gain deeper insight into the properties of optical neural networks.

System performance has always been limited and will always be limited by device performance. The development of devices with improved performance depends also on the demand for commercial systems which make use of them. The liquid crystal television panels are a good example for that. The addressing of individual pixels had been a problem for years. With the upcoming of multimedia and portable computing, the demand for discretely addressable panels was high enough to stimulate the development of such devices. Such developments are of course more related to markets, economics and research funding than to physics and engineering. In the context of this work, the question is whether liquid crystal light valves will become a reliable mass product or not. The demand for this could come from a similar domain as for the LCTVs: A large screen television projector system that bases on LCLVs is currently being developed by Gretag. A success of this technology on the market would probably increase the availability of reliable LCLVs. Alternatively, the demand could come from optical processing systems like, for example, optical neural networks.

At this point of the discussion we have to look back to the remarks that have been made about optical neural networks in the introduction to this thesis. Currently, optical neural networks are not competitive with alternative solutions, i.e. software and electronic solutions. There are particular advantages of the optical approach, related to the parallel processing power of optics, however, what counts in reality is the ensemble of numerous factors like performance, cost, size, reliability, maintainability, flexibility, etc. There are two conclusion for future work:

1) The neural network aspect is given up and one concentrates on optical parallel processing modules that are universally usable. An example would be the matrix-vector-multiplier. The presented concept can be miniaturized by the integration of VCSEL technology and specialized opto-mechanics. The interlaced fan-out method offers an elegant way to realize a large number of parallel channels with minimal imaging errors. The use of a dilute input array is interesting in the context of the heat related problems of VCSELs.

2) Neural network applications that cannot satisfactorily be solved by other approaches have to be exploited. One promising approach has been proposed recently by Farhat [47, 48]. He describes a class of neural networks, where clusters of smaller netlets are nonlinearly coupled. Such networks should be able to process information that cannot be handled by conventional neural network models, especially dynamic input patterns (spatio-temporal signals). Our setup would be suitable for the optical implementation of such a network, if the LCLV is replaced by a smart pixel structure (as described above), where every smart pixel represents a neural netlet.

Acknowledgments

Many people contributed directly or indirectly to the successful completion of this thesis. I'd like to thank

- Prof. R. Dändliker for giving me the opportunity to make a thesis in his group, for his continuous support of my work and for his many helpful comments and advises.
- The members of the jury - M. Gale, Prof. H. Hügli and Prof. J. Jahns - for their critical review of my thesis.
- The Swiss National Science Foundation for the financial support of this project.
- A. Pourzand for being an almost unlimited source of all kind of practical information, for many fruitful discussions and for the great time we had together during the four years that we shared our office.
- Dr. R. Völkel for his introduction to "Raytrace", for his many hints concerning optical systems and for his numerous critical but very helpful remarks to my work.
- Dr. N. Collings for introducing me to the field of optical information processing and for the personal contacts and opto-electronic hardware that he organized for me.
- M. Gale, Dr. Th. Hessler and Dr. M. Rossi of the CSEM Zürich for the fabrication of numerous diffractive microlens arrays. Their generous and uncomplicated support was an essential contribution to this work.
- Dr. K. Weible for the fabrication of various diffractive elements and for helpful hints about fabrication technology.
- Ph. Nussbaum and I. Philipoussis for the fabrication of various refractive microlens arrays.
- D. Ruffieux and Dr. K.-H. Gulden of the CSEM Zürich for the fabrication of the VCSEL-array and for the information about the state of the art of VCSEL technology.
- Prof. P. Seitz and Dr. O. Vietze of the CSEM Zürich for lending me a detector array, and Dr. J. Piot and A. Sasseli of Logitech SA for providing the detector electronics.
- J.-L. Kumin and J.-R. von Allmen of the IMT Neuchâtel for their superb mechanical constructions.
- M. Groccia for his support in all questions of electronics.
- P. Blattner for many valuable hints about programming with MATLAB and for his help with the literature collection.

- J.M. Teijido for introducing an inveterate Mac-lover to the mysterious world of PC configuration.
- My semester students T. Jost and D. Gehriger for their preparatory work with the LCLV.
- The administrative and technical staff of IMT Neuchâtel for their invaluable contribution to the infrastructure of the institute.
- All the colleagues of the Applied Optics Group for their daily support, for the numerous apéros, for the common swims in the nearby lake, for their help with the French language and for the friendly atmosphere.
- My neighbors, H. Schmocker and F. Lerch, for lending me a study room in the basement of our house.
- My parents-in-law for giving asylum to my family in the final days of the thesis.
- My wife Barbara and my children for their continuous love, support and patience during the last months.
- My parents. Words will never be sufficient to express my gratitude for all that they have done for me.

Thank you!

Appendix A A short introduction to neural networks

The potential and limitations of artificial neural networks and the difficulties in modeling human information processing can only be understood if one has an idea about the complexity of the processes in our brain. An overview of the structure of the human brain and its mechanisms of information processing is therefore given in section A.1. The general characteristics of artificial neural networks are presented together with a chronologically ordered list of different network models in section A.2. In section A.3, the Hopfield-model is explained in more detail, because it has been the base for our (and many other) optical implementation of neural networks.

A.1 Biological neural networks

The theory of neural networks evolved from the efforts to understand the mechanisms of information processing in the human brain. To understand the structure of artificial neural networks, it is indispensable to have some basic knowledge about the structure of the human brain.

A good part of the information in this section is taken from a special edition about the brain of "Spektrum der Wissenschaft" (German edition of "Scientific American"), N° 11/1992. [1, 49-51]

A.1.1 The human brain and the cortex

The human brain consists of several elements with different tasks (Fig. A.1). The essential part of human information processing happens in the outer part of the brain, the *cortex*. The cortex is a soft tissue, about 2-3 mm thick with an area of ~0.2 m^2. The cortex is again divided into regions with different functionality. It contains around 10^{11} cells called *neurons*, which are the basic processing elements.

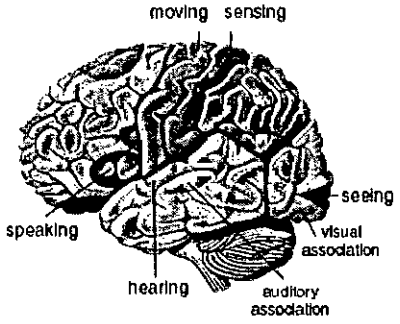


Fig. A.1: Side view of the cerebral cortex of the human brain. Different regions are responsible for different task.

A.1.2 The neurons

Numerous forms of neurons have been observed by means of various coloring techniques (Fig. A.2). The basic structure, however, is the same for all neurons: Each neuron consists of a cell body, the *soma*, around 10'000 "receiver antennas", the *dendrites*, and one "sender antenna", the *axon*. The axon can be rather long (up to 1 m) and ramifies at the end.

Fig. A.2: Photo of neurons, made visible with the so-called Golgi coloring technique. One can recognize the cell bodies (soma) with one long sender element (axon, above cell body) and many smaller receiving elements (dendrites, below cell body). (From [1])

The many dendrites of a neuron receive signals in the form of electrical pulses from other neurons (Fig. A.3). The soma sums the incoming signals. When sufficient input is received, the cell fires, i.e. it transmits a signal over its axon to other cells. The

electrical pulses of ~100 mV amplitude and ~1 ms duration are generated by a mechanism of different Na-ion concentration inside and outside of the neuron cell and a variable permeability of the cell membrane. The frequency of pulses that a cell can fire is limited to ~200 pulses/second due to the chemical processes involved. The signal propagation speed along the axon (~100 m/s) is more than 10^6 times slower than the propagation of an electrical signal within a copper wire. Both frequency and signal propagation are thus much slower than in electronic computer systems. The superiority of our brain for some tasks like face or voice recognition is due to the massively parallel operation of many neurons. The total number of interconnections between neurons in the cortex is in the order of 10^{14}.

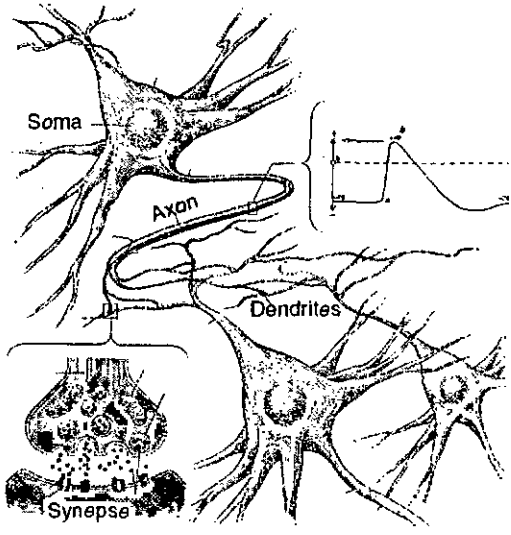


Fig. A.3: Signals are exchanged between neurons by means of electrical pulses, generated by the cell body (soma). At the interface between the sender element (axon) of one neuron and the receiver element (dendrite) of another neuron is the synapse, where the signal is excited or inhibited by electrochemical processes (blow-up lower left). (From [1])

A.1.3 The synepses

The interface between the axon of one neuron and the dendrite of another neuron is called *synapse*. The transmission of the electrical signals across a ~20 nm wide gap happens by means of chemical substances, the neurotransmitters (Fig. A.3, blow-up lower left). These transmitters can be either stimulating or inhibitory. The signal

transmission efficiency of a synapse can vary over time, by chemical changes either on the sender side (more transmitters produced) or on the receiver side (receptors more sensitive). Increased transmission efficiency can be initiated by simultaneous activity of the two neurons before and after the synapse [52]. It is generally believed today that the process of learning is related to changes in signal transmission efficiency of the synapses in the cortex.

A.1.4 Learning

At the moment of birth, the cortex already contains the basically complete set of neuron cells. However, these neurons are hardly interconnected. There is a rough interconnection pattern (whose structure is genetically coded), but the fine-tuning of the growth of axons and dendrites happens due to external stimulation during the first years of life.

As mentioned above, the process of learning is related to changes in the signal transmission efficiency between neurons. There are basically two ways to accomplish this: by chemical changes in the synapses or by changes in the anatomic structure of the interconnections, i.e. the growth of new axon endings and the steady change of the dendrite-trees. The latter has been demonstrated e.g. by Merzenich *et al.* [51] and by Hawkins *et al.* [53]. Merzenich measured the size of cerebral regions that were activated by the movement of the fingers of a monkey. After some months of intense training of only three of the monkeys fingers, the corresponding regions were substantially larger, partially on the cost of the other finger regions.

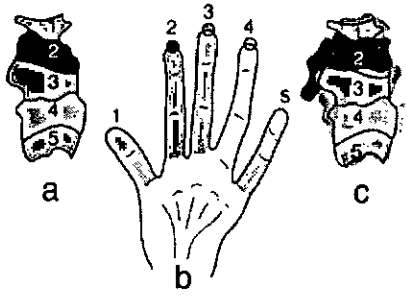


Fig. A.4: Schematic sketch of the cerebral regions (a) that correspond to the fingers of the left hand of a rhesus monkey (b). After a few months of intense training of fingers 2, 3 and 4, the corresponding regions have grown, partially on the cost of the other finger regions (c). (From [51])

Hawkins *et al.* performed extensive experiments with Aplysia californica, a variety of a sea slug. Because Aplysia's brain has only about 20'000 neurons they have been

able to chart which groups of neurons were producing neurotransmitters in response to an external stimulus, and they could also observe the growth of new interconnections between neuron groups.

The regrouping or growth of interconnections is obviously related to long-term memory. Chemically induced changes of synaptic efficiency, on the other hand, were observed for short-term and for long-term memory.

A.1.5 Human information processing and storage

The interconnection patterns in the human brain are very complex, and a lot is still unknown today. The goal of this subsection is not to give a complete overview but to show a few examples in order to illustrate the complexity of human information processing.

The current knowledge of the mechanisms in the human brain bases on various techniques. By means of electrodes or using tomographic methods, it is possible to measure the activity of different regions of the brain in function of a specific stimulation (Fig. A.5).

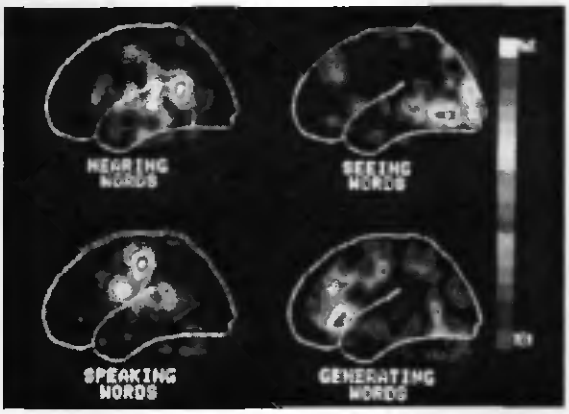


Fig. A.5: Activity of different brain regions for different actions, measured by a tomographic method. Top left: hearing words; top right: seeing words; lower left: speaking words; lower right: generating words.(From [1])

The spatial and temporal resolution of such methods is still rather poor, but using ingenious tests it has been possible to obtain a lot of interesting results. It has been shown, for example, that the recognition of form, color and movement of an object happens in different regions of the visual cortex [50, 54]. It has also been shown that

other neurons are responsible for the processing of lines than for the processing of spots.

Another way to gain insight in the mechanisms of human information processing is to study the mental capabilities of patients with partially damaged brains. One interesting example has been reported by Moscovitch *et al.* at the University of Toronto [55]. Their patient, C.K., lost his ability to read and to recognize objects after a car accident with head injuries. However, C.K. was still able to recognize faces. He could even recognize modified faces and caricatures of famous persons. Experiments were also made with images of the Italian painter Giuseppe Arcimboldo whose paintings show faces composed of vegetables, flowers and other objects (Fig. A.6). Though C.K. was not able to see the individual objects, he could easily recognize the faces.

Fig. A.6: An example for the complexity of human information processing: An image of the Italian painter Guiseppe Archimboldo has been shown to a patient with a partially damaged brain. While the patient was still able to recognize the face, he was not capable anymore to realize that the face is composed of vegetables and flowers.

Various other studies on patients with partially damaged brains [56] showed that different kind of information is stored in many different ways and places in our brain. Some patients lost the ability to recall events from the time before their accident, but they could easily recall events after the accident. Remarkably, some information from the past remained intact. One patient, for example, could not remember that he once learnt Italian, but he still understood the language. Others showed just the inverse behavior. They still knew everything about their past life but they were not able to memorize events or persons they met after their accident [57]. Accidents with locally

isolated injuries caused the loss of very specific abilities, e.g. the ability of seeing moving objects, whereas all other visual functions remained intact. Other examples include the loss of short-term memory with intact long-term memory and vice versa.

Stepping forward now to artificial *neural* networks, one has to be aware of the huge parallelism and the very complex nature of human information processing, and of our still poor knowledge about it in order to avoid exaggerated expectations.

A.2 Artificial neural networks

A.2.1 General characteristics

As their biological models, artificial neural networks process information by means of many simple processing units, which are called *neurons*, *units*, *cells* or *nodes*. Basically, the neurons are binary elements, i.e. they are active (they "fire") or they are inactive. The neurons pass signals (which correspond to their state) between each other over directed *connection links*. Each connection link has an associated *weight*, which is typically used to multiply the transmitted signal. Each neuron applies a nonlinear function (called *activation function*, *transfer function* or *threshold function*) to the sum of its (weighted) inputs.

An artificial neural network is characterized by 1) the pattern of connections between the neurons (called the *architecture* of the neural net), 2) the method of determining the weights on the connections (*training* or *learning algorithm*), and 3) the threshold function of the neurons (which is normally identical for all the neurons of a net).

The information that a neural net can use to solve a problem is stored in the ensemble of all its weights, which is normally represented by a *weight-matrix*. The states of all neurons with identical function (e.g. all input neurons) are represented by corresponding vectors (e.g. input vector). Sets of neurons with identical functionality are called *layers* (e.g. input layer).

A.2.2 History and mathematical models

The history of artificial neural networks goes back to 1943, when McCulloch and Pitts [58] introduced the artificial neuron as a logical thresholding element with several inputs and one binary output. Several neurons could be combined to implement various logical functions. Their nets had fixed weights and were not able to learn.

In 1949, Hebb [52] designed the first learning law for artificial neural networks. He postulated that the strength of a synapse between two neurons should be increased when both of the neurons were active at the same time. First computer simulations showed that this rather general statement had to be refined.

1958, Rosenblatt [59] presented a new class of neural networks, the perceptron. The typical perceptron consists of an input layer (e.g. the retina) connected to a layer of associator neurons. The weights of the connections can be adjusted using a learning rule that adjusts the weights of a unit whenever the response of the unit is incorrect.

Closely related to the perceptron and its learning rule is the work of Widrow and Hoff [60] who introduced the so-called delta-rule, which adjusts the weights to reduce the difference between the net output for a given input and the desired net output. This (slight) improvement to the learning rule leads to an improved ability of the net to generalize.

In 1969, the development of artificial neural networks was slowed down by the book of Minsky and Papert [61] who proved that perceptrons and related types of neural nets are not able to implement certain classes of logical functions, e.g. the XOR-function.

In 1982, Hopfield [3] introduced a new type of neural nets with feedback that has been the base for many future neural network models, especially for pattern recognition.

In 1986, Rumelhart, Hinton and Williams [62] presented the back-propagation algorithm that allowed to train multi-layer networks. Such networks do not fall under the restrictions of the single layer perceptron (as described by Minsky and Papert, cf. above). The back-propagation algorithm is the base for a lot of today's neural network applications.

Further information about artificial neural networks can be found e.g. in [6], [63] and [64]. An extensive overview of various hardware implementations and applications is given e.g. in [65]. A good starting point for information about neural network hardware on the world wide web is a page at CERN [66].

A.3 The Hopfield model

The Hopfield model will be described now in more detail, because it has been the base for our and many other optical implementation of a neural network.

A.3.1 Description

A Hopfield net consists of one single layer of N neurons. Every neuron acts as both input and output. Each neuron can have two possible states (0 and 1) and is connected with every other neuron of the net. The states of all the neurons are represented by a vector $\vec{n} = (n_1, n_2, ..., n_N)$. The connection between neuron n_m and neuron n_n has an associated weight W_{nm}. The ensemble of all weights is stored in the weight matrix W. After initialization with an input vector $\vec{n}_{[0]}$, the network iterates through multiple states according to the equation

$$\vec{n}_{[i+1]} = f(W \cdot \vec{n}_{[i]}) \quad , \tag{A-1}$$

where i is the iteration number and f is the threshold function. If the input vector is similar to one of the reference vectors that has been coded in the weight matrix (cf. Eq. A-4), the network will converge towards a stable state after a certain number of iterations. The stable state of the network corresponds to the (recognized or restored) reference vector. The Hopfield type of neural networks can thus be used as an associative memory, e.g. for pattern recognition.

In the original Hopfield model, the non-linear threshold function is a binary function

$$f(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \le 0 \end{cases} . \tag{A-2}$$

The calculation of the weight matrix is made according to the following rule:

$$W_{mn} = \begin{cases} \sum_{k=1}^{K} (2r_m^{[k]} - 1)(2r_n^{[k]} - 1) & \text{if } m \neq n \\ 0 & \text{if } m = n \end{cases} , \tag{A-3}$$

where K is the number of reference vectors $\vec{r}$ and $r_m^{[k]}$ is the state of the m -th element of the k -th reference vector $\vec{r}^{[k]}$. The contribution of one reference vector $\vec{r}^{[k]}$ to W_{mn} is 1 if $r_m^{[k]} = r_n^{[k]}$ and -1 if $r_m^{[k]} \neq r_n^{[k]}$.

The update of the neuron states according to Eq. (A-1) has to be asynchronous, i.e. the neurons have to be updated randomly but at the same average rate. This and the

zero weights on the diagonal of the weight matrix are important features that guarantee the stability and convergence of the net [6, pp. 137-140].

The convergence of the network can be analyzed by introducing an energy function, in analogy to the Hamilton function in phsical systems. In the energy landscape which is defined by this function, the minima correspond to the possible stable states of the system.

The storage capacity of such a neural network is limited to $0.15 \cdot N$ reference vectors (N = number of neurons) [3].

A.3.2 Modified Hopfield models

Numerous variations exist for the Hopfield type of neural networks. Here, some models relevant for optical implementations of the Hopfield model shall be presented. For details, the reader is referred to the literature.

e) Continuous model

Hopfield himself [5] adapted his model to allow for a continuous threshold function like the sigmoid-function

$$f(x) = \frac{1}{1+e^{-cx}} \quad . \tag{A-4}$$

Hopfield showed that the continuous model shows similar behavior than the discrete model if the weight matrix is symmetric and the threshold function is sufficiently steep. Also, the input activation was allowed to continue after the first iteration in this modification.

The possibility to have a continuous threshold function is important for the use of a device like the LCLV as a thresholding element.

b) Model with bias

The motivation of this and the following models is the difficulty to optically realize bipolar weights.

Jang, Jung, Lee and Shin [67] introduced an additional constant value to the interconnects in order to make them unipolar. Such biasing, however, involves dynamic threshold levels for the neurons, which is difficult to implement optically.

c) Inhibitory model

Shariv and Friesem [45] showed that the performance of the Hopfield network remains essentially unchanged when all the positive weights in Eq. A-4 are set to zero.

As a drawback of this method, each neuron obtains a variable threshold. This can be compensated for by normalizing the weight matrix and by adapting the threshold function. The number of patterns that can be stored in such a network is reduced to $0.11 \cdot N$ [68].

d) Inverted model

Shariv, Grossman, Domany and Friesem [46] introduced a model that is based on the inhibitory model, but uses a binary weight matrix, which is an advantage for the optical implementation (easier fabrication, more throughput). The memory capacity is slightly reduced from $0.11 \cdot N$ to $0.10 \cdot N$, as a result of the sacrificed dynamic range [68].

A.3.3 Advantages and limitations

The computational power of the Hopfield neural network lies in many very simple identical logical units (the neurons). Hopfield networks offer the possibility of concurrent distributed processing that enables a massively parallel structure without the need for synchronization.

Hopfield networks are relatively insensitive to local imperfections (e.g. variations in the threshold level of individual neurons or of the weights in the interconnections), which makes them suitable for real-world hardware implementations.

On the other hand, the number of patterns that can be stored is limited (smaller than 0.15 times the number of neurons, as mentioned above). When the number of neurons is increased, the number of interconnections increases with the square. The requirements for the practical implementation of a larger network increase correspondingly.

The patterns (or vectors) stored in the weight matrix should not share too many common pixels (mathematically this is described as a large Hamming-distance), otherwise the net tends to converge to local minima or to show metastable states [69].

Appendix B Some optics background

For convenience, some basics of optics that have been used during the work are summarized here. For detailed information, the reader is referred to the literature, e.g. [15, 16, 21, 70-72].

B.1 F-number and numerical aperture

The **f-number** of a lens with focal length f and circular aperture of diameter D (Fig. B.1) is defined by

$$f_\# = \frac{f}{D} \quad . \tag{B-1}$$

Related to this is the **numerical aperture**

$$NA = n\sin(\alpha) \quad , \tag{B-2}$$

where n is the refractive index of the surrounding medium and α the half-angle of the illumination cone of the lens. For a source placed in the front focal plane and for small angles, the relation between the two terms is given by

$$f_\# = \frac{1}{2 \cdot NA} \quad . \tag{B-3}$$

The numerical aperture is a measure for the amount of light emerging from a source that can be captured by a specific lens. A lens with a large aperture (small f-number) is said to be "fast" and a lens with a small aperture (large f-number) is said to be "slow". The terminology comes from photography, where lenses with large aperture (i.e. a lot of light can be captured) allow for the use of fast films or short illumination times.

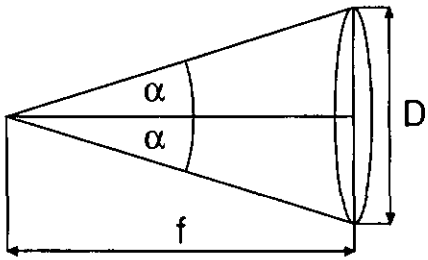


Fig. B.1: Numerical aperture and f-number are used to quantify the acceptance angle of a lens (cf. text).

B.2 Diffraction limited spot size of a focused beam

A detailed discussion of the topic can be found e.g. in [15, chapter 10, 16, section 4.3].

A laser beam is focused using a lens of focal length f with a circular aperture of diameter D. Diffraction limits the size of the focal spot to a minimal diameter d of

$$d = 2.44 \frac{\lambda f}{D} \quad . \tag{B-4}$$

83.9 % of the total beam energy are within this diameter, the rest is in outer rings.

In the case of a rectangular aperture of width D_x and D_y, the focal spot has a rectangular shape of width d_x and d_y, which is given by

$$d_x = 2 \frac{\lambda f}{D_x} \quad \text{and} \quad d_y = 2 \frac{\lambda f}{D_y} \quad . \tag{B-5}$$

These formulas represent a *theoretical minimum* for an ideal optical system. In a *real* optical system with imperfections and aberrations, the achievable spot size may be larger.

B.3 Aberrations

Commonly used optics formulae are only exactly valid for an infinitesimal region around the optical axis (paraxial rays). Aberrations are all kind of deviations of off-axis rays from the paraxial image points. They are present in all real-world optical systems with finite aperture and finite field of view.

Excellent introductions to the topic can be found e.g. in [15, section 6.3, 21, chapter 3]. Here, only the basic terms are summarized.

The primary aberrations (or Seidel aberrations) of a system in monochromatic light are spherical aberration, coma, Petzval curvature, astigmatism and distortion. Chromatic aberrations, i.e. deviations of the paraxial image points due to the variation of the index of refraction with wavelength, will not be covered here.

Spherical aberration can be defined as the variation of focus with aperture. As a ray is moved towards the edge of a lens (farther away from the paraxial region), the position of the ray intersection with the optical axis moves farther away from the paraxial focus. In general, for a positive (converging) lens, the intersection will move closer to the lens (undercorrected spherical aberration). For a given aperture and focal length, the amount of spherical aberration is a function of object position and shape of the lens. The image of a point formed by a lens with spherical aberration is usually a bright dot surrounded by a halo of light. The effect of spherical aberration to an extended image is to soften the contrast and to blur the details.

Coma can be defined as the variation of magnification with aperture. When a bundle of oblique rays is incident on a lens with coma, the rays passing through the edge portions of the lens are imaged at a different height than those passing through the center portion. Coma varies with the shape of the lens and also with the position of any apertures which limit the bundle of rays that form the image. The image of a point formed by a comatic lens looks like a raindrop or (thus the name of this aberration) a comet.

Petzval curvature is a sort of basic field curvature associated with every optical system. It is depending on the index of refraction of the lens elements and on their surface curvature. The effect of field curvature is that the image is not laying on a plane but rather on a parabolic surface. Positive (converging) lenses produce a Petzval surface that is curved towards the lens (inward, or undercorrected). For a thin, simple element, the longitudinal deviation is proportional to the square of the image height and inversely proportional to the focal length and the refractive index of the lens. Petzval curvature can be reduced or avoided by combining suitable positive and negative lenses.

Astigmatism comes from the fact that a bundle of rays can hit the lens under different angles in the meridional (or tangential) plane and in the sagittal plane [15, p. 240]. From Fermat's principle follows that the effective focal length of the lens gets shorter if the rays are more oblique. This results in a different focal position of the meridional and the sagittal rays. The image of a point formed with a astigmatic lens consists of two lines; a horizontal line formed by the meridional rays and a vertical line formed by the sagittal rays. Between these lines, the image is an elliptic or

circular blur. The image of an extended object will be formed on two parabolic surfaces, the sagittal and the meridional focal surfaces. Astigmatism is related to the Petzval curvature: When there is no astigmatism, the sagittal and meridional focal surfaces coincide with each other and lie on the Petzval surface.

Distortion comes from the fact that the transversal magnification can be a function of the distance of the object point from the optical axes. The image of a square object will look like a pincushion if the distortion is positive (or overcorrected) and it will look like a barrel if the distortion is negative (or undercorrected).

B.4 The Gaussian beam

The principles of geometrical ray optics base on idealized light sources. An ideal plane wave, for example, is not possible in the real world because it would have to extend over the whole infinite space. An ideal spherical wave cannot exist either, because it assumes a source of ideal point size where the energy would be infinite.

Laser beams produced by single-mode lasers can be well described by Gaussian beams. The intensity distribution of such beams in any transverse plane is a circularly symmetric Gaussian function, i.e. the energy is concentrated in a small cylindrical region along the beam axis. The width w of these beams is minimum at the beam waist ($z = 0$; $w = w_o$) and grows gradually in both directions. The wavefronts are planar at the waist, but they gradually become curved, approximately spherical, while propagating.

A detailed description of this subject can e.g. be found in [16, chapter 3]. In the following, some properties relevant for the understanding of the effects described in chapter 2 shall be discussed.

B.4.1 Properties

A Gaussian beam propagates according to the equation

$$w(z) = w_0\left[1+\left(\frac{z}{z_0}\right)^2\right]^{1/2}, \qquad \text{(B-6)}$$

where $w(z)$ is the beam radius at the longitudinal position z. The radius is defined as the lateral distance from the beam axis, where the intensity I dropped to $1/e^2$ of its maximum value I_o. 86.5 % of the beam power is encompassed within the beam

radius. The radius w_o at $z = 0$ is called the *waist radius*, and z_0 is the *Rayleigh length*, which is defined by

$$z_0 = \frac{\pi}{\lambda} w_0^2 \quad , \tag{B-7}$$

thus

$$w_0 = \left(\frac{\lambda z_0}{\pi}\right)^{1/2} \quad . \tag{B-8}$$

The waist diameter $2w_o$ is called the *spot size*. The radius $w(z)$ of the Gaussian intensity distribution increases with increasing $|z|$. At $z = z_0$, it becomes $\sqrt{2}w_0$. The axial distance within which the beam radius lies within this factor is called the *depth of focus*. The depth of focus is thus equal to $2z_0$. For $z \gg z_0$, Eq. (B-6) becomes

$$w(z) \approx \frac{w_0}{z_0} z = \Theta_0 z \quad , \tag{B-9}$$

where Θ_0 is called the *beam divergence*. Using Eq. (B-7), the beam divergence can be expressed as

$$\Theta_0 = \frac{\lambda}{\pi w_0} \quad . \tag{B-10}$$

For a constant wavelength, the beam divergence is thus inversely proportional to the waist radius.

Due to the spread of the beam radius, the intensity on the beam axis decreases with increasing $|z|$. At a distance $z = \pm z_0$, the intensity drops to 1/2 of its value at $z = 0$ according to

$$I(z) = I_0 \left[\frac{w_0}{w(z)}\right]^2 = \frac{I_0}{1 + (z/z_0)^2} \quad . \tag{B-11}$$

The *radius of curvature* $R(z)$ is given by

$$R(z) = z\left[1 + \left(\frac{z_0}{z}\right)^2\right] \quad . \tag{B-12}$$

From this formula, one can see that the Gaussian beam can be approximated as a plane wave around the waist and as a spherical wave far from the waist. The radius of curvature is smallest at $z = z_0$.

B.4.2 Pessage through optical componants

The *passage of a Gaussian beam through a thin lens* of focal length f (circular lens, aligned with the beam axis) can be described by

$$\frac{1}{R'} = \frac{1}{R} - \frac{1}{f} \quad . \tag{B-13}$$

The Gaussian beam remains a Gaussian beam after passage through a thin lens. Only the curvature and the position of the waist are changed, i.e. the beam is only reshaped by the lens. By introducing a magnification

$$M = \frac{\left|\frac{f}{z-f}\right|}{\sqrt{1+\left(\frac{z_0}{z-f}\right)^2}} \quad , \tag{B-14}$$

where z is the distance of the lens from the front side waist, it can be shown that the parameters of the reshaped beam are given by

$$w_0' = M w_0 \qquad \text{(waist radius)}, \tag{B-15}$$

$$(z'-f) = M^2(z-f) \qquad \text{(waist location)}, \tag{B-16}$$

$$2z_0' = M^2(2z_0) \qquad \text{(depth of focus)}, \tag{B-17}$$

$$\Theta_0' = \frac{\Theta_0}{M} \qquad \text{(beam divergence)}. \tag{B-18}$$

In the limit where $(z - f) \gg z_0$ (lens far away from the waist of the incident beam), M can be approximated by

$$M \approx \left|\frac{f}{z-f}\right| \quad . \tag{B-19}$$

Applying this approximation to Eq. (B-16) leads to the imaging equation of ray optics:

$$\frac{1}{f} = \frac{1}{z} + \frac{1}{z'} \quad . \tag{B-20}$$

Howevar, if a lens is placed with its front focal plane coinciding with the waist of the incident beam, $f = z$, thus $M = 1$. Therefore the waist of the outgoing beam will be in

the back focal plane ($z' = z = f$). From geometrical optics one would have expected collimation from such a configuration!

A Gaussian beam is *collimated* by making the location of the new waist as distant as possible from the lens. This is achieved by using the smallest ratio z_0/f. For a given ratio z_0/f, the optimal value of z for collimation is $z = f + z_0$.

Relevant in the context of this work is the fact, that true collimation does not exist in reality. After a certain distance, the beam will always spread. When using arrays of microlens-telescopes with a "collimated" region within two arrays, we have to keep the distance between the arrays small enough to avoid crosstalk between neighboring channels.

If the depth of focus of the incident beam $2z_0$ is much longer than the focal length f of a lens (as it is the case for a lens placed in the raw beam of a laser), $M = f/z_0$ and $z' = f$. That means that the transmitted beam is focused at the back focal plane of the lens as one would expect for a plane wave. Indeed, the Gaussian beam can be approximated by a plane wave in the surrounding of its waist. The spot size is given by

$$2w_0' = 2\frac{f}{z_0}w_0 = 2\frac{\lambda f}{\pi w_0} \quad . \tag{B-21}$$

The effect of truncating a Gaussian beam by an aperture is discussed by several authors [73-76]. If the diameter of the aperture is not reduced below $2\cdot(2w_0)$, then the beam intensity distribution remains within a few percent of a Gaussian distribution. Below this value, an aperture will introduce rings into the irradiance pattern and the result approaches that of a plane wave focused by a lens. In practice, the lens diameter will rather be $2w_0$, which introduces an energy loss of 13 % and increases the effect of diffraction spreading [77].

B.5 Diffraction grating and pitch of generated spot-array

A diffraction grating consists of a repeating structure of periodicity Λ. When illuminated with a collimated beam of wavelength λ, the beam is divided in several beams of different diffraction order n. The angles of the diffracted beams depend on the periodicity and on the wavelength of the laser beam, whereas the *intensity distribution* within the diffraction orders depends on the nature of the repeating structure. The n-th diffraction order is deviated by an angle α_n, which is given by

$$\sin\alpha_n = n\frac{\lambda}{\Lambda} \quad . \tag{B-22}$$

The height Δ_n above the optical axes of this beam at a given distance d is given by

$$\tan\alpha_n = \frac{\Delta_n}{d} \quad , \tag{B-23}$$

thus

$$\Delta_n = d \cdot \tan\left(\sin^{-1}\left(n\frac{\lambda}{\Lambda}\right)\right) \quad . \tag{B-24}$$

For $\Lambda \gg \lambda$ (small angles) this can be simplified to

$$\Delta_n = d \cdot n\frac{\lambda}{\Lambda} \quad . \tag{B-25}$$

When the grating is used in the front focal region of a lens, d has to be replaced by the focal length f of the lens. Δ_n is equal to the pitch of the generated spots with $n = 1$ for standard gratings and $n = 2$ for gratings with an "even-order-missing"-design.

Appendix C Refractive and diffractive microlens arrays

Our setup and its successful implementation depended heavily on microlens arrays, as shown in chapter 2. We tested, compared and used various types of microlenses. In this appendix, different types of microlens arrays shall be compared with respect to their use in our setup.

The discussion of the microlens arrays is divided in diffractive and refractive elements. Further information can be found e.g. in [78].

C.1 Diffractive microlens arrays

C.1.1 General properties of diffractive lenses

The potential of diffractive lenses (and diffractive elements in general) is strongly related to their fabrication methods that allow an extremely high degree of freedom for the design of the element. Asymmetric aspheric lenses, arrays of lenses with spatially variant properties, etc. are examples of optical elements which are difficult or impossible to fabricate with traditional methods. Large diffractive lenses are flatter and lighter than their refractive counterparts, which makes them suited for portable applications. As for arrays of lenses, unconventional lens shapes (rectangles, hexagons, etc.) are possible without loss of the optical quality in the edges. This allows for an array fill factor of 100 %. The focal length of diffractive lenses can be exactly predetermined by the geometrical structure of the lens. The quality of the profile will influence the efficiency of the lens but not the focal length (for large Fresnel numbers, cf. [79]). This property is also an advantage in the context of replication. Compared to their refractive counterparts, a larger range of f-numbers is feasible (cf. below). For a given lens size, the lower limit for the f-number (≥ 1) is determined by the minimal feature size at the edges of the lens [36]. Moving towards the upper limit results in a transition to the refractive case.

Potential problems of diffractive lenses include the strong dependence of the focal length of the wavelength and smaller efficiency due to diffraction losses (see e.g. [23]). The wavelength dependence can eventually be an advantage when the lens is used in combination with a diffraction grating because the lateral chromatic

dispersion of the two elements compensate each other. However, the longitudinal changes remain [80, 81].

C.1.2 Fabrication and releted properties

Extensive overviews of different fabrication methods can be found in [82] for binary elements and [83] for continuous relief elements.

Here, we only want to mention the two methods that have been employed for the fabrication of the elements used in this work: direct laser beam writing [35, 36] and photolithography. Both methods provide excellent precision of the lateral geometry, which is important when the arrays are used in a demanding configuration like ours, with large f-numbers and low tolerances for the lateral displacement of individual focal spots. The fabrication of large master elements is rather time consuming and expensive for both methods. The advantage of the continuous relief is an increased efficiency compared to the multilevel binary elements. The flexibility of the laser beam writing method is a huge advantage during the construction of a prototype setup that uses many different elements. On the other hand, continuous relief elements can suffer from problems related to the scanning process (finite size of the writing spot and superposed grating structure caused by stage positioning errors).

C.1.3 Replication

In order to lower the cost and to allow mass production, several replication techniques for diffractive elements have been developed. An extensive overview of replication techniques can be found in [31].

The basic principle is to fabricate a negative (nickel shim) of the original structure by electroforming. From this negative, high-resolution replicas can be made by hot embossing (stamping or rolling), UV casting or embossing, and injection molding (as for the CD fabrication). Each technique has its special advantages like cost, speed, resolution, etc. The common materials for the replicas are polycarbonate, PMMA (embossing and molding) and epoxy materials (curing).

All replication technologies are capable of very high fidelity reproduction of sub-micrometer sized features and some of them (hot roller embossing and injection molding) are well established industrial technologies. Additional surface roughness introduced by the replication is negligible.

A potential problem for certain applications can be the limited planarity of large Ni shims due to stress introduced during the galvanoform process (see also below). Recent optimizations of the galvanoform process helped to significantly reduce this problem [33]. An alternative to the use of a Ni shim is to transfer the original

photoresist structure into a hard material, e.g. fused quartz, by reactive ion etching. The different etch speeds and the different refractive indices of photoresist and quartz have to be respected during the design of the original.

UV curable materials have a lateral shrinkage of about 1 % on curing. A shrinkage in the same order of magnitude can also be observed for the hot embossing process. This shrinkage can in principle be compensated for during the design and fabrication of the original.

System designers have to keep some special properties of replicas in mind: Some plastic materials are birefringent, which can be a problem in a polarization sensitive setup. The thermal expansion coefficient of plastic is about 10 times larger than the one of glass, which may be of concern under real-world conditions.

C.1.4 Experimentel work

e) Continuous relief microlens arrays (PSI/CSEM Zürich)

Most of the diffractive microlens arrays used throughout this project are continuous relief Fresnel elements, fabricated in photoresist with the laser beam writing system of the CSEM Zürich (former PSI Zürich) [35, 36].

As mentioned in section 2.2.1 (boundary conditions), the lateral geometry of our microlenses was determined by the pitch of the LCTV pixels.

Early elements, used in the comparison of the two fan-out principles (section 2.5), had 16 x 16 microlenses with 25 mm focal length (at a design wavelength of 633 nm) and a pitch of 1400 µm x 1150 µm (corresponding to 25 pixel pitches of 56 µm x 46 µm of our old EPSON VPJ-2000 LCTVs). The lateral size of the lenses was equal to the pitch, the fill factor was thus 100 %. Fig. C.1.a shows one lens with its neighbors and Fig. C.1.b shows a blow-up of the region between two lenses. The horizontal grating of ~2 µm pitch that is superposed to the lens structure originates from the finite spot size of the writing laser.

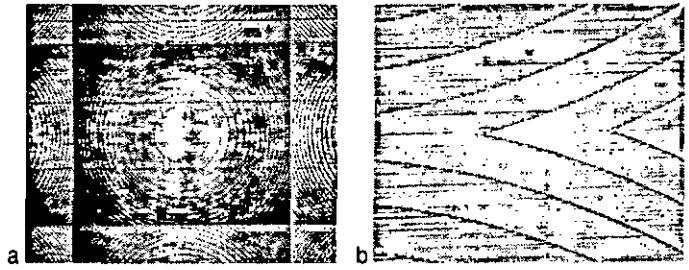


Fig. C.1: a) Photo of a continuous relief microlens array from the early testing phase. b) In the right photo we can see a superposed horizontal grating that originates from the finite spot size of the laser writing process.

The main series of experiments was made with various arrays of 16 x 16 microlenses with 11.54 mm, 50.00 mm and 80.00 mm focal length respectively (at a design wavelength of 488 nm) and a pitch of 1092 µm x 1092 µm (corresponding to 26 pixel pitches of 42 µm x 42 µm of our new InFocus LitePro 850 LCTVs). The lateral size of the lenses was equal to the pitch, the fill factor thus again 100 %. Additional details about these elements are provided in [33, pp. 64-66].

Measured efficiencies were between 55 % (early elements) and 75 % (variations are due to different profile qualities). As mentioned above, elements with continuous relief offer in principle a higher efficiency than corresponding binary elements. Surface roughness due to the finite spot size (cf. Fig. C.1.b) and small stage positioning errors during the scanning process caused losses (scattered light) in the order of a few percent [33]. In addition, an imperfection in the current version of the absolute positioning system of the scanning table caused a superposed grating of 8 µm period (cf. Fig. C.2.a). We measured that ~10.6 % of the incident light was diffracted into higher orders of this grating. Finally, the photoresist is not completely transparent at 488 nm, which gave rise to further losses. For all these reasons, the efficiency of the continuous relief elements was reduced to the level of multilevel elements (cf. below).

The performance of individual lenses was diffraction limited, i.e. the measured spot size corresponds to the diffraction limited spot sizes.

Fig. C.2.a shows a blow-up of one lens out of an array of 50 mm-microlenses directly written into photoresist. The sample has been illuminated in order to show the superposed grating structure. Fig. C.2.b shows the same for an element that has been transferred into fused quartz (see below). The profile of the superposed grating has been flattened during the transfer. Fig. C.2.c and Fig. C.2.d show the diffraction limited spots generated by illuminating the lenses of Fig. C.2.a and Fig. C.2.b respectively. The bright spots at the upper and lower edge of Fig. C.2.c are the first diffraction orders of the superposed grating. The influence of this grating is clearly reduced after the transfer of the element into fused quartz (Fig. C.2.d).

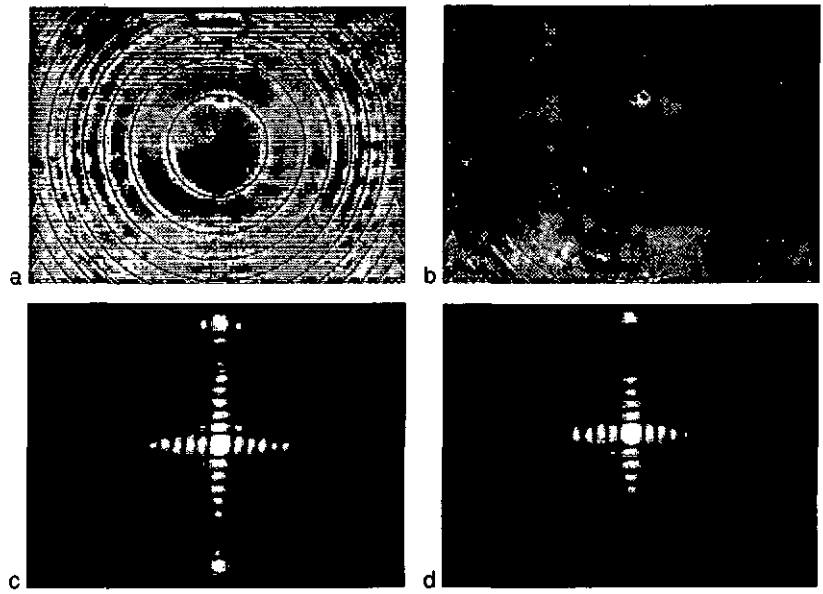


Fig. C.2: a) Blow-up of one 50 mm lens directly written in photoresist. The sample has been illuminated to show the superposed grating, originating from the fabrication. b) Dito for an element that has been transferred into fused quartz. The superposed grating has been flattened out during the replicetion. c) Photo of a spot generated by illuminating one original lens (as shown in a). The spot size is diffraction limited. The large spots close to the upper end lower edge are caused by the superposed grating. d) Dito for the replicated element. The spot size is unchanged end the influence of the superposed grating is considerably reduced.

b) Replices

The use of replicas has been considered for our setup, because of the large number of microlens arrays required. 4 replicas in polycarbonate foil, a UV-casted replica and one sample that has been transferred into fused quartz by reactive ion etching have been tested and compared. All elements were made from laser-beam written originals with continuous relief and 50 mm focal length.

The *quality of the single lenses* did not suffer from the replication process, which is in accordance with the remarks in section C.1.1. All replicas still showed diffraction limited spot sizes (Fig. C.2.d) and efficiencies of ~60 %. The element that had been transferred into fused quartz showed the highest efficiency of ~80 %. This is probably due to the better transparence of quartz compared to photoresist and to a smoothing out of the superposed grating (Fig. C.2.b and Fig. C.2.d). On the other hand, the amount of scattered light was increased for this element.

To test the *uniformity of the replicated arrays*, we used a setup corresponding to the input generation method of section 2.3.1. We illuminated them by a plane wave and compared the generated spot array to a reference pattern on the LCTV that was placed in the focal plane. This method is similar to the Shack-Hartmann wavefront sensor, but with a plane reference-wavefront and an unknown test-lens-array. Whereas the spot arrays generated by the photoresist originals correctly matched the reference pattern (Fig. C.3.a), the polycarbonate and the UV-casted replicas showed considerable deviations (up to 120 µm, Fig. C.3.b). Only the element that has been transferred into fused quartz showed the same flatness as the originals. The deviations were related to slight stress-induced distortions of the Ni-shim during the galvanic electroform process. Whereas this is not a problem for the replication of small elements and gratings, it is one in a demanding application like ours with large array size, long focal lengths and low tolerance for lateral errors. A lateral displacement of 10 µm in the focal plane (z = 50 mm) corresponds to a 200 nm tilt over one lens aperture of ~1 mm. According to [33], this replication process has been significantly improved in the meantime.

In addition to the problems related to the Ni-shim flatness, the elements replicated in polycarbonate foil suffered from a shrink of ~0.74 % due to the hot embossing process. In principle, this shrink can be compensated for during the fabrication of the original element. However, we observed that the deviations resulting from the shrink were different for every sample.

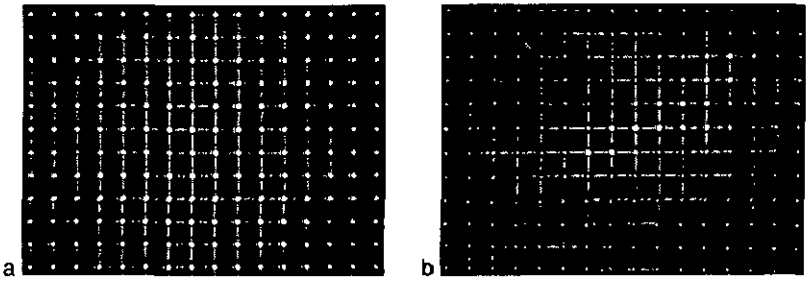


Fig. C.3: A spot array is generated by illuminating a microlens array by a plane wave. The array is compared with a reference pattern on an LCTV placed in the focal plane. a) Perfect match for the photoresist original; b) mismatch caused by insufficient planarity for the replicas (here: UV-casted replica).

As a consequence of these tests, only original elements directly written into photoresist have been used in the final setup (cf. Table 2.2). They satisfied our demanding requirements.

c) Multilevel elements (Weible OpTech, Neuchâtel)

Two pairs of Fresnel microlens arrays have been fabricated with classic photolithography methods by Weible OpTech, Neuchâtel. One pair had a focal length

of 11.54 mm and the other had a focal length of 80 mm. The lateral geometry was identical to the continuous relief elements mentioned above. The measured efficiency of the 4-level elements was ~75 % (80 mm-elements) and ~63 % (11.53 mm-elements). The elements satisfied our requirements and were successfully used in the final setup (cf. Table 2.2).

The tested multilevel and continuous-relief elements performed equally well. After improvements in fabrication technology (elimination of superposed grating, transfer into a more transparent material), the continuous-relief elements will be more efficient.

C.2 Refractive microlens arrays

C.2.1 General properties of refractive lenses

Compared to their diffractive counterparts, refractive lenses do not have diffraction losses. The efficiency of a refractive lens with antireflection coatings can be as high as ~99 %. Wherever losses are an issue, refractive lenses should be considered. Another plus of refractive lenses is the low dependence of their focal length on the wavelength. This can be important in combination with a source or an array of sources with unstable or non-uniform wavelength such as laser-diodes and VCSELs.

The design freedom for the fabrication of refractive lenses, however, is much smaller, because the fabrication of suitable refractive surface profiles is not a trivial task. This is true for macro-optics and micro-optics, but from now on this discussion is limited to arrays of refractive microlenses.

C.2.2 Fabrication and related properties

A common method to fabricate arrays of refractive microlenses is to melt flat structures (e.g. cylinders) of photoresist [84, 85]. Surface tension is responsible for the reflow of the material, e.g. to a hemispherical shape. This and other techniques including photothermal techniques and the swelling of irradiated PMMA are described in [86] and [87].

With the melting resist technology, it is possible to manufacture lenses with apertures of all kind of shapes [88]. The resulting surface profiles, however, are a result of surface tensions and are not directly controllable. Obviously, discontinuities are not possible. In consequence, the surface profile at the edges of unconventional shapes is normally not optically ideal. For arrays of refractive microlenses this leads to an upper limit of the fill factor. Arrays with circular apertures have better optical quality but the fill factor is smaller than $\pi/4 = 78.5$ % (square arrangement) or $\pi/2\sqrt{3} \approx 90.7$ % (hexagonal arrangement). Arrays of rectangular or hexagonal lenses allow for a higher fill factor at the cost of inferior optical quality at the edges of the individual lenses. A minimal gap between individual lenses has to be respected in any case in order to allow for the forming of the desired surfaces.

The difficulty to achieve a good uniform coating of thick (~100 µm) resist, the contact angle of the melted resist with the substrate surface and gravity (for large lenses) are factors that limit the range of possible lens dimensions.

In practice, lenses with acceptable quality can be made with f-numbers ranging from 1 to 5. In order to fabricate lenses with higher f-numbers, pre-shaping techniques [89] or an index match liquid with $n_{liquid} < n_{resist}$ and a glass cover can be used.

In contrast to comparable diffractive lenses, refractive lenses are rather thick. This is important when photoresist originals are used at wavelengths where the absorption of the photoresist is high.

C.2.3 Replication

The replication methods for refractive microlens arrays and the related properties are in principle the same as in the diffractive case. The potentially larger thickness can limit the practical feasibility, e.g. in the case of the transfer of a thick photoresist original into a fused quartz substrate.

C.2.4 Experimentel work

Several arrays of refractive microlenses have been tested. The elements have been fabricated at IMT Neuchâtel with the melting resist technology. The arrays had 16 x 16 lenses with a pitch of 1092 µm x 1092 µm. The lens aperture was circular with a diameter slightly smaller than the pitch. Because of the limited range of achievable lens dimensions (cf. above), the focal length was only 5.5 mm. The height (maximal thickness) of the photoresist lenses was ~60 µm.

After initial problems (deviation of the correct lens position and focal length variations, caused by a non-uniform base layer), elements with good lateral geometry have been fabricated. They were tested in a similar arrangement as the diffractive elements (cf. section C.1.4). The lateral position of the generated focal spots satisfied our requirements. However, the focal length was too short for a use of these elements in the optical neural network (cf. section 2.2.2).

The thickness of the photoresist lenses results in a high absorption, which lowered the efficiency to ~38 %. A transfer of the microlens-array into fused quartz would improve this value, but the thickness of the sample (~60 µm) is at the limit of the currently feasible for the transfer technique.

To increase the transparency, a replicated element made by UV-casting into an elastomeric mold [90] has been tested. To increase the focal length, an index matching liquid has been put between the lenses and a thin glass substrate. The focal length could be increased by a factor of ~5, but the shrink of the element during the replication (cf. section C.1.3) made it useless for our setup. No experiments have been made with index matching liquid in combination with original photoresist lenses, because of the too low efficiency of the latter.

The technique of melting photoresist is very attractive for arrays of small microlenses (diameter ≤ 500 µm). The fabrication of relatively large lenses (~1 mm diameter, as desired in our setup) of good quality is possible, but the absorption of the photoresist of such large lenses reduces the efficiency below an acceptable value. Technological progress might allow the transfer of such elements into fused quartz in the future. The limited range of possible f-numbers can be increased by using index matching liquid. No refractive microlens arrays were used in the final version of our setup.

C.3 Comparison and conclusion

Table C.1 shows a summary of properties of diffractive and refractive microlens arrays. The consequences for the use of such elements in a system like ours are given in the third column.

Type / Property	Diffractive	Refractive	General remarks and consequences for use in our specific setup
Efficiency	good	in principle better; in our tests reduced by absorption of photoresist	Use refractive lenses when efficiency is an issue. A transparent material (e.g. fused quartz) is needed to profit from this advantage.
Dependence of focal length on wavelength	strong	weak	Important in the combination with laser diodes or VCSELs.
Fill factor	100 %	< 100 %	LA3 and LA4 should be diffractive elements, because the whole square aperture is needed.
Aperture shapes	arbitrary; good optical prop. also in corners	arbitrary; good optical prop. only for circular aperture or in center	
Predicteble focal length	yes	no precise prediction today; can be precisely measured	Important in combination with a diffraction grating (LA3): deviations in the focal length result in deviations of the spot array pitch.
f-number	≥ 1	< 5, expandable by index matching liquid	In our system with relatively thick (20 mm) commercial mounts and a lens diameter of ~1 mm, diffractive elements are used.

Table C.1: Comparison of some properties of diffractive and refractive microlens arrays and the resulting consequences for the use of such elements in our setup.

Appendix D Alignment of the optical setup

This appendix has been added with regard to the eventual use of the presented optical system or a modified version of it by someone else in the future.

The following text is not meant to fully describe every detail. After several months of intense work with an optical setup, one gets a feeling for its behavior which is hard to write down in words. But hopefully these lines show some basic concepts and help to facilitate the work with the setup.

D.1 Preparatory work

This will certainly sound trivial, but it is important enough to be mentioned anyway: A well prepared beam is the unalterable base for any successful system work. Enough time and care should therefore be spent for this part of the work. The use of a shear-plate to test the collimation of the beam and of a penta-prism to get orthogonal beams helps to considerably speed up the process. Pre-alignment of the single system components and the use of one single reference target also helps to save a lot of time.

D.1.1 Beam expansion

Beam expansion is achieved by using a microscope objective of focal length f_1 and an achromat of focal length f_2. Microscope objectives are normally classified by their magnification M (10x, 20x, 40x, ...). f_1 can roughly be estimated by

$$f_1 = \frac{\text{tubus length}}{M} \quad , \tag{D-1}$$

where the standard tubus length is 160 mm. The exact value for f_1 is normally slightly different and can be found in the manufacturer catalogs.

The ratio between the raw laser beam diameter D_1 and the expanded beam diameter D_2 is calculated by using the formulas of Appendix B.4.2 (Gaussian beam).

The achromat is oriented with its flat side towards the objective in order to minimize spherical aberrations. Its z-position for best collimation of the expanded beam is adjusted by using a shear-plate collimation tester.

A Spindler & Hoyer 10X-objective and a S & H 100 mm achromat (Ø 50 mm) have been used in this setup for beam expansion.

D.1.2 Spatial filtering

In order to get a clean beam, a pinhole is placed in the focal plane of the microscope objective. The dimension of the pinhole can be determined as follows: The beam waist w_1 in the focal plane of the microscope objective is given by

$$w_1 = \frac{\lambda f_1}{\pi w_0} \quad , \tag{D-2}$$

where w_0 is the waist of the raw laser beam. $2w_0 \approx 1.22$ mm for our Spectra Physics Ar^+-laser. The diameter of the pinhole should now be $D_{ph} \geq 2w_1$. A detailed description of the effects of apertures of various size to a Gaussian beam can be found in [76].

A 15 µm pinhole has been used in this work.

D.2 Alignment

D.2.1 General concept

In this subsection, the *general concept* that was used for the alignment of a spot array to an aperture array is presented.

The basic principle behind the method of aligning an array of spots to an array of apertures (detector elements, LCTV pixels, smart pixels, etc.) is the Moiré–effect: If one array is rotated (around the z-axis) with respect to the other, a fringe pattern can be observed (Fig. D.1). The larger the tilt between the arrays, the smaller the period of the fringes. If the aperture array is moved horizontally, the fringes move vertically and vice versa.

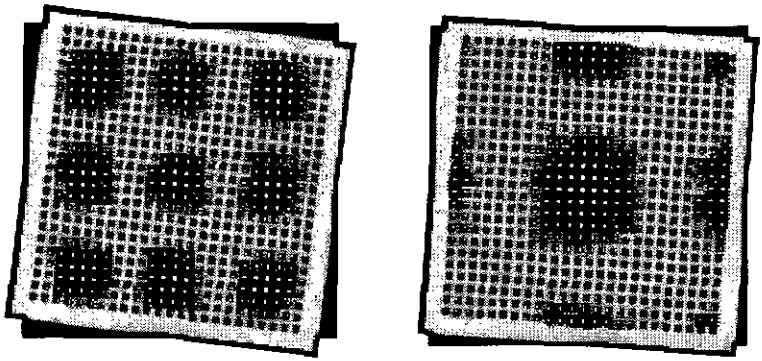


Fig. D.1: Moiré-fringes are used to align an array of spots to an array of apertures. The smaller the fringes, the larger the tilt between the two arrays. If the aperture is moved horizontally, the fringes move vertically and vice versa.

After rough adjustment of ϑ_z, the alignment is tested by slightly displacing the aperture array (LCTV) by half a pixel pitch. If all the spots disappear simultaneously, the arrays are well aligned. If they do not, there is still a slight tilt, a magnification error (different pitch of spot array and aperture array) or an imaging error. The kind of observed fringe-movement helps to isolate the source of the error.

The sensitivity to a lateral displacement of the aperture array is an indication to the local spot size and thus to the z-alignment. The aperture-array is in the focal plane when the sensitivity is highest.

The Moiré-effect can also be used to magnify small structures [91]. This has been used for the z-alignment of the collimating microlens arrays. The microlens array is slightly rotated around the z-axis. The sharpness of the so produced magnified spot pattern is observed at a large distance (laboratory wall) and is an indication for the z-position of the array.

For some of the alignment steps, the described patterns can be observed on a screen by eye. In some cases a CCD camera is used together with reducing optics. The use of a complete subsystem (camera and optics mounted together on an OWIS System-40 rail which can be mounted on the OWIS System-65 rail) proved to be very helpful. This allows to easily observe different planes with always the same image size.

D.2.2 Pre-alignment of components

All the beamsplitters and mirrors are previously aligned to a precise 90° horizontal deviation in the expanded and collimated beam. The reference deviation is determined by means of a penta-prism.

The initial spot array (generated by Dammann-grating DG1 and lens L0, cf. e.g. Fig. 2.26) is aligned to a carefully centered reference target. The x- and y-positions are adjusted by x- and y-movement of L0, ϑ_z is adjusted by rotating the DG1. This step is done for both the input and the feedback part of the system.

Both LCTVs are then pre-aligned with respect to the initial spot array. A sequence of alignment patterns (crossed lines → spot array with helping lines → spot array) is displayed on the LCTV to facilitate the approach to the correct position. The method described in section D.2.1 (slight movement of LCTV) is used to check the correct position.

Finally, the microlens arrays are pre-aligned. A test pattern (frame with diagonals) is displayed on LCTV1 and the microlens array is used to collimate the beams emerging from LCTV1. The collimated beams are aligned with respect to the reference target by adjusting the x-, y- and ϑ_z-position of the microlens-array.

D.2.3 Alignment of the initialization part

The precise alignment of LCTV1 and LA2a is made identically to the pre-alignment described above.

Thanks to the combination of interlaced fan-out and microchannel-telescopes the beamsplitter BS2 does not influence the (collimated) beams emerging from LA2 (perpendicular incidence). Moreover, the following optical elements (DG2, LA3, LCTV2) are to some extent independent of the previous elements. For example, the z-distance of this group of elements from LA2a can be chosen freely, as long as the effect of beam spreading is not too large (cf. section 2.4.2 and Appendix B.4).

The Dammann-grating DG2 is placed at the front focal plane of LA3. It was found that a manual measurement of the distance between DG2 and LA3 is sufficient. Slight deviations of this distance result in a slight deviation from the perpendicular incidence on LCTV2. The position of the spots on LCTV2, however, is not influenced. The ϑ_z-position of DG2 can normally remain unchanged if the pre-alignment has been made carefully.

The z-position of LCTV2 is aligned next. LA3 is slightly rotated in order to produce a Moiré-pattern together with LCTV2. Due to the small depth of focus in this image plane, the correct position can be found by looking for the largest contrast of the Moiré-image (i.e. some beams pass through a pixel and some beams are blocked in the region between the pixels).

Due to the non-ideal holder of LCTV2 (the degrees of freedom are not independent), the x- and y-position of LCTV2 are not changed anymore after pre-alignment (this is actually the hardest element to pre-align). Instead, the x-, y- and ϑ_z-position of LA3 is slightly changed until the spot array matches the pixel array. The kind of used test-patterns can be seen in Fig. 2.32. As mentioned above (D.2.1), the alignment is checked by slight x- and y-movement of LA3. The alignment is correct if the whole signal emerging from LCTV2 disappears simultaneously.

D.2.4 Alignment of the feedback part

It is important that the feedback optics is aligned before the fan-in optics. The fan-in optics should only be added when both the fanned-out image from the input and the fanned-out image from the feedback properly match on LCTV2.

In order to generate an alignment test-pattern (frame with diagonals), a mask is positioned at the focal plane of L0b (equivalent position of LCTV1). LA2b is aligned identically to LA2a with the reference target.

The same remarks as above concerning the beamsplitter PBS3 and the z-position of LA6 can be made.

The LCVL z-, ϑ_x- and ϑ_y-position are aligned to maximize the reflected signal at the (s-polarized, i.e. deviated) output of PBS3.

Finally, the signal reflected from the LCLV has to be matched with the test pattern on LCTV2 by alignment of the two beamsplitters PBS3 and BS2. At the same time, the system output at BS2 has to match for both initialization and feedback mode.

D.2.5 Alignment of the fan-in part

LA4 is placed at a distance from LCTV2 which roughly corresponds to its focal length. L5 is placed a few centimeters behind LA4. M2 is a flip mirror which is set to non-deviating for observation of the matrix-vector-multiplier output. A transparent screen (chalk paper) is used to observe the image that is formed by LA4 and L5. The z-position and ϑ_z of LA4 have to be aligned first. They are good if it is possible to observe a magnified image of the LCTV pixels (one pixel ~ 0.6 mm x 0.8 mm). The x- and y-position of LA4 are aligned by positioning the matrix-vector-multiplier output with respect to the reference target.

Finally, the two mirrors M2 and M3 are aligned to match the matrix-vector-multiplier output incident on the write side of the LCLV with the spot array incident on the read side of the LCLV. Correct alignment is achieved when the test pattern (frame with diagonals) can be modulated to the LCLV output.

References

[1] G. D. Fischbach, *Gehirn und Geist*, in "Spektrum der Wissenschaft", vol. 11/1992, pp. 30-41 (1992).

[2] D. Psaltis and N. Farhat, *Optical information processing based on an associative-memory model of neural nets with thresholding and feedback*, Opt. Lett. **10** (2), pp. 98-100 (1985).

[3] J. J. Hopfield, *Neural networks and physical systems with emergent computational abilities*, Proc. Natl. Acad. Sci. USA **79**, pp. 2554-2558 (1982).

[4] J. W. Goodman, A. R. Dias, and L. M. Woody, *Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms*, Opt. Lett. **2** (1), pp. 1-3 (1978).

[5] J. J. Hopfield, *Neurons with graded response have collective computational properties like those of two-state neurons*, Proc. Natl. Acad. Sci. USA **81**, pp. 3088-3092 (1984).

[6] L. Fausset, "Fundamentals of neural networks", 1st ed., Prentice-Hall Inc., Englewood Cliffs, NJ (1994).

[7] S. Jutamulia, *Selected papers on optical neural networks*, in "SPIE milestone series", SPIE, Bellingham/Washington (1994).

[8] C. Denz, "Optical neural networks: an introduction with special emphasize on photorefractive implementations", 1st ed., Vieweg, Braunschweig/Wiesbaden (1998).

[9] K. J. Weible, *Experimental investigation of optical neural networks and learning systems*, Ph.D. thesis, University of Neuchâtel, Neuchâtel (1993).

[10] W. Xue, *Characterization of liquid crystal light valves for neural network applications*, Ph.D. thesis, University of Neuchâtel, Neuchâtel (1994).

[11] A. R. Pourzand, *Optimization of 2D liquid crystal devices for use in optical information processing systems*, Ph.D. thesis, University of Neuchâtel, Neuchâtel (1998).

[12] C. M. Marcus and R. M. Westervelt, *Basins of attraction for electronic neural networks*, presented at IEEE Conference on Neural Information Processing Systems, Denver, CO (1987).

[13] L. S. Lee, H. M. Stoll, and M. C. Tackitt, *Continuous-time optical neural network associative memory*, Opt. Lett. **14**, pp. 162-164 (1989).

[14] J. R. Leger, *Laser Beam Shaping*, in "Micro-Optics", pp. 223-257, H. P. Herzig, Ed., Taylor & Francis Ltd., London/Bristol (1997).

[15] E. Hecht, "Optik", Addison-Wesley (1989).

[16] B. E. A. Saleh and M. C. Teich, "Fundamentals of Photonics", John Wiley & Sons, Inc., New York (1991).

[17] H. Dammann and K. Görtler, *High-efficiency in-line multiple imaging by means of multiple phase holograms*, Opt. Comm. **3** (5), pp. 312-315 (1971).

[18] H. Dammann and E. Klotz, *Coherent optical generation and inspection of two-dimensional periodic structures*, Optica Acta **24** (4), pp. 505-515 (1977).

[19] M. P. Dames et al., *Efficient optical elements to generate intensity weighted spot arrays: design and fabrication*, Appl. Opt. **30**, pp. 2685-2691 (1991).

[20] A. Vasara et al., *Binary surface-relief gratings for array illumination in digital optics*, Appl. Opt. **31**, pp. 3320-3336 (1992).

[21] W. J. Smith, "Modern Optical Engineering", 2nd ed., Mc Graw-Hill (1990).

[22] N. Lindlein, *Raytrace*, University of Erlangen-Nürnberg, Erlangen (1994).

[23] H. P. Herzig, *Design of refractive and diffractive micro-optics*, in "Micro-Optics", pp. 1-29, H. P. Herzig, Ed., Taylor & Francis Ltd., London/Bristol (1997).

[24] T. Hessler, *Personal Communication*, CSEM, Zürich (1997).

[25] Yoshikawa et al., IEEE J. of Quantum Electron. **34** (6), pp. 1009 ff. (1998).

[26] K. Gulden, "Report to the EU about the future development of VCSELs," CSEM, Zürich (1998).

[27] T. J. Cloonan, *Architectural considerations for optical computing and photonic switching*, in "Optical Computing Hardware", pp. 1-43, J. Jahns and S. H. Lee, Eds., Academic Press Inc., San Diego (1994).

[28] A. W. Lohmann, *Scaling laws for lens systems*, Appl. Opt. **28** (3), pp. 4996-4998 (1989).

[29] A. W. Lohmann, *Image formation of dilute arrays for optical information processing*, Opt. Comm. **86**, pp. 365-370 (1991).

[30] F. B. McCormick, F. A. P. Tooley, T. J. Cloonan, J. M. Sasian, H. S. Hinton, K. O. Mersereau, and A. Y. Feldblum, *Optical interconnections using microlens arrays*, Opt. Quantum Electron. **24**, pp. 465-477 (1992).

[31] M. T. Gale, *Replication*, in "Micro-Optics", pp. 153-177, H. P. Herzig, Ed., Taylor & Francis Ltd., London/Bristol (1997).

[32] T. Hessler, M. Rossi, J. Pedersen, M. T. Gale, M. Wegner, D. Steudle, and H. J. Tiziani, *Microlens arrays with spatial variation of the optical functions*, JEOS A, Pure and Appl. Opt. 6, pp. 673-681 (1997).

[33] T. Hessler, *Continuous-relief diffractive optical elements*, Ph.D. thesis, University of Neuchâtel, Neuchâtel (1997).

[34] M. C. Hutley, P. Savander, and M. Schrader, *The use of microlenses for making spatially variant optical interconnections*, JEOS A, Pure and Appl. Opt. **1**, pp. 337-342 (1992).

[35] M. T. Gale and K. Knop, *The fabrication of fine lens arrays by laser beam writing*, Proc. SPIE **398**, pp. 347-353 (1983).

[36] M. T. Gale, M. Rossi, J. Pedersen, and H. Schütz, *Fabrication of continuous-relief micro-optical elements by direct laser writing in photoresist*, Opt. Eng. **33** (11), pp. 3556-3566 (1994).

[37] N. Collings and C. Berger, *Demonstration and discussion of an interlaced fan-out interconnect*, Inst. Phys. Conf. Ser. **139** (Part II), pp. 247-250 (1994).

[38] C. Berger, N. Collings, and T. Jost, *Recurrent optical neural network for the study of pattern dynamics*, OSA Technical Digest Series 8, pp. 46-48 (1997).

[39] A. A. Vasiliev, I. N. Kompanets, and A. V. Parvenov, *Advances in development and applications of optically controlled liquid crystal spatial light modulators*, Optik 67 (3), pp. 223-236 (1984).

[40] J. W. Goodman, "Introduction to Fourier optics", 2nd ed., McGraw-Hill (1996).

[41] P.-G. de Gennes, "The Physics of Liquid Crystals", Clarendon Press, Oxford (1974).

[42] Y. D. Dumarevskii, N. F. Kovtonyuk, I. N. Kompanets, and A. V. Parfenov, *Metal-insulator-semiconductor-liquid crystal structures. Influence of parameters of control signals on characteristics of spatial modulation of light*, Sov. J. Quantum Electron. **14** (4), pp. 493-496 (1984).

[43] R. Dändliker, *Optique Appliquée I & II*, Lecture Notes EPFL, Lausanne (1994).

[44] H. J. White, *Experimental results from an optical implementation of a simple neural network*, Proc. SPIE **963**, pp. 570-575 (1988).

[45] I. Shariv and A. A. Friesem, *All-optical neural network with inhibitory neurons*, Opt. Lett. **14**, pp. 485-487 (1989).

[46] I. Shariv, T. Grossman, E. Domany, and A. A. Friesem, *All-optical implementation of the inverted neural network model*, Proc. SPIE **1319**, pp. 194-195 (1990).

[47] N. H. Farhat, *Dynamical networks with bifurcation processing elements*, presented at the 1997 Symposium on Nonlinear Theory and Application, Hawaii (1997).

[48] N. H. Farhat, *Dynamical computing with diverse attractors*, Proc. OII'98, J. of Optoelectron. **9 Supp.**, pp. 137-139 (1998).

[49] C. J. Shatz, *Das sich entwickelnde Gehim*, in "Spektrum der Wissenschaft", vol. 11/1992, pp. 44-52 (1992).

[50] S. M. Zeki, *Das geistige Abbild der Welt*, in "Spektrum der Wissenschaft", vol. 11/1992, pp. 54-63 (1992).

[51] E. R. Kandel and R. D. Hawkins, *Molekulare Grundlagen des Lernens*, in "Spektrum der Wissenschaft", vol. 11/1992, pp. 66-76 (1992).

[52] D. O. Hebb, "The Organization of Behavior", John Wiley & Sons, New York (1949).

[53] R. D. Hawkins, T. W. Abrams, T. J. Carew, and E. R. Kandel, *A cellular mechanism of classical conditioning in Aplysia*, Science **219**, pp. 400-404 (1983).

[54] S. M. Zeki, *Functional specialization in the visual cortex of the rhesus monkey*, Nature 274 (5670), pp. 423-428 (1978).

[55] M. Moscovitch, G. Winocur, and M. Behrmann, *What is special about face recognition?: Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition*, J. of Cognitive Neuroscience 9, pp. 555-604 (1997).

[56] L. Johnson and J. Neffe, *Die Schatten der Erkenntnis*, in "GEO Wissen - Intelligenz und Bewusstsein", pp. 28-39, Gruner + Jahr, Hamburg (1992).

[57] B. Milner, *Amnesia following the operation on the temporal lobes*, in "Amnesia: clinical, psychological and medicolegal aspects", C. W. M. Whitty and O. L. Zangwill, Eds., Butterworths (1966).

[58] W. S. McCulloch and W. Pitts, *A Logical calculus of the ideas immanent in nervous activity*, Bulletin of Mathematical Biophysics 5, pp. 115-133 (1943).

[59] F. Rosenblatt, *The perceptron: A probabilistic model for information storage and organization in the brain*, Psych. Rev. 65 (6), pp. 386-408 (1958).

[60] B. Widrow and M. Huff, *Adaptive switching circuits*, IRE WESCON Convention Record part 4, pp. 96-104 (1960).

[61] M. L. Minsky and S. A. Papert, "Perceptrons", MIT Press, Cambridge, MA (1969).

[62] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagation error*, Nature 323, pp. 533-536 (1986).

[63] J. A. Anderson and E. Rosenfeld, *Neurocomputing: Foundations of Research*, . Cambridge, London: MIT Press (1988).

[64] W. Kinnebrock, "Neuronale Netze", 2nd ed., Oldenbourg (1994).

[65] E. Sánchez-Sinencio and C. Lau, *Artificial neural networks: paradigms, applications, and hardware implementations*, in "A Selected Reprint Volume", vol. PC0279-0, IEEE, New York, pp. 336 tt. (1992).

[66] http://www1.cern.ch/NeuralNets/nnwInHepHard.html

[67] J.-S. Jang, S.-W. Jung, S.-Y. Lee, and S.-Y. Shin, *Optical implementation of the Hopfield model for two-dimensional associative memory*, Opt. Lett. 13, pp. 248-250 (1988).

[68] H. Shouval, I. Shariv, T. Grossman, A. A. Friesem, and E. Domany, *An all-optical Hopfield network: Theory and experiment*, Int. J. of Neural Systems 1, pp. 355-360 (1991).

[69] R. P. Lippmann, *An Introduction to Computing with Neural Nets*, in "IEEE ASSP Magazine", pp. 4-22 (1987).

[70] H. Haferkorn, "Optik", 3rd ed., Barth Verlagsgesellschaft, Leipzig/Berlin/ Heidelberg (1994).

[71] H. Kuchling, "Physik", 16th ed., VEB Fachbuchverlag, Leipzig (1983).

[72] M. Young, "Optik, Laser, Wellenleiter", Springer, Berlin/Heidelberg/New York (1997).

[73] J. P. Campbell and L. G. DeShazer, *Near fields of truncated-Gaussian apertures*, J. Opt. Soc. Am. **59** (11), pp. 1427-1429 (1969).

[74] L. D. Dickson, *Characteristics of a propagating Gaussian beam*, Appl. Opt. 9 (8), pp. 1854-1861 (1970).

[75] R. G. Schell and G. Tyras, *Irradiance from an aperture with a truncated-Gaussian field distribution*, J. Opt. Soc. Am. **61** (1), pp. 31-35 (1971).

[76] P. Belland and J. P. Crenn, *Changes in the characteristics of a Gaussian beam weakly diffracted by a circular aperture*, Appl. Opt. **21** (3), pp. 522-527 (1982).

[77] F. Sauer, J. Jahns, C. R. Nijander, A. Y. Feldblum, and W. P. Townsend, *Refractive-diffractive micro-optics for permutation interconnects*, Opt. Eng. 33 (5), pp. 1550-1560 (1994).

[78] H. P. Herzig, *Micro-Optics*, Talyor & Francis Ltd., London/Bristol (1997).

[79] T. Hessler and R. E. Kunz, *Relaxed fabrication tolerances for low-Fresnel-number lenses*, J. Opt. Soc. Am. A 14 (7), pp. 1599-1606 (1997).

[80] M. Schwab, N. Lindlein, J. Schwider, Y. Amitai, A. A. Friesem, and S. Reinhorn, *Achromatic diffractive fan-out systems*, in "Diffractive and Holographic Optics Technology", vol. 2152, pp. 14-20, I. Cindrich and S. H. Lee, Eds., SPIE, Bellingham (1994).

[81] G. P. Behrmann and J. N. Mait, *Hybrid (refractive / diffractive) optics*, in "Micro-Optics", pp. 259-292, H. P. Herzig, Ed., Taylor & Francis Ltd., London/Bristol (1997).

[82] M. B. Stern, *Binary optics fabrication*, in "Micro-Optics", pp. 53-85, H. P. Herzig, Ed., Taylor & Francis Ltd., London/Bristol (1997).

[83] M. T. Gale, *Direct writing of continuous-relief micro-optics*, in "Micro-Optics", pp. 87-126, H. P. Herzig, Ed., Taylor & Francis Ltd., London/Bristol (1997).

[84] Z. D. Popovic, R. A. Sprague, and G. A. N. Connell, *Technique for monolithic fabrication of microlens arrays*, Appl. Opt. **27** (7), pp. 1281-1284 (1988).

[85] D. Daly, R. F. Stevens, M. C. Hutley, and N. Davies, *The manufacture of microlenses by melting photoresist*, J. Meas. Sci. Techn. **1**, pp. 759-766 (1990).

[86] M. C. Hutley, *Refractive lenslet arrays*, in "Micro-Optics", pp. 127-152, H. P. Herzig, Ed., Taylor & Francis Ltd., London/Bristol (1997).

[87] M. Kufner and S. Kufner, "Micro-optics and Lithography", VUBPRESS, Brussels (1997).

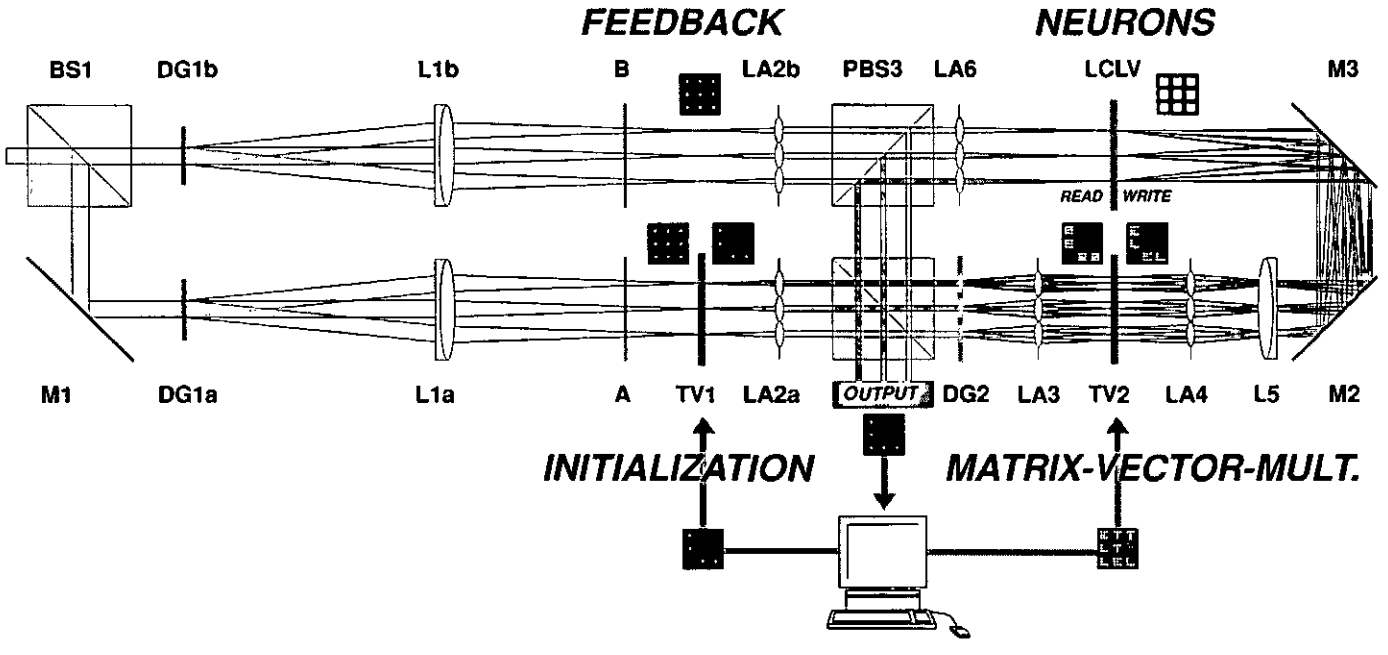
[88] R. Völkel, P. Nussbaum, K. J. Weible, H. P. Herzig, R. Dändliker, S. Haselbeck, M. Eisner, and J. Schwider, *Fabrication of non-conventional microlens arrays*, EOS Topical Meetings Digest Series 5, pp. 116-120 (1995).

[89] T. R. Jay and M. B. Stern, *Preshaping photoresist for refractive microlens fabrication*, Opt. Eng. **33** (11), pp. 3552-3555 (1994).

[90] P. Nussbaum, I. Philipoussis, A. Husser, and H. P. Herzig, *Simple technique for replication of micro-optical elements*, Opt. Eng. **37** (8), pp. 1804-1808 (1998).

[91] M. C. Hutley, R. Hunt, R. F. Stevens, and P. Savander, *The Moiré Magnifier*, Pure Appl. Opt. 3, pp. 133-142 (1994).

Compact All-Optical Recurrent Neural Network



Scheme of the optical setup