PAPER TO BE PRESENTED AT THE FOURTH INTERNATIONAL CONFERENCE ON ARTIFICIAL NEURAL NETWORKS - ANN 95

CAMBRIDGE, UK, 26-28 JUNE 1995.

INVARIANT OBJECT RECOGNITION WITH A NEUROBIOLOGICAL SLANT

J. Keat, V. Balendran, K. Sivayoganathan, and A. Sackfield*.
Manufacturing Automation Research Group,
Department of Manufacturing Engineering,
*Department of Mathematics and Statistics,
The Nottingham Trent University,
Burton Street,
Nottingham, NG1 4BU,
UK.

tel: (0115) 985 6075 ex 2349
e-mail : keat@lister.ntu.ac.uk

INVARIANT OBJECT RECOGNITION WITH A NEUROBIOLOGICAL SLANT

J. Keat, V. Balendran, K. Sivayoganathan,
Student Member, MIMechE, MIEE, UK.

The panacea of object recognition has to be a general solution that can be applied to all problems, but as yet this is just a distant dream. The ability of even the youngest child to recognise and perceive an object irrespective of it's view shows how far object recognition has still to go. Marr (1) describes this as being 'primarily (a) phenomena of information processing', and he stated that if this problem is to be addressed it must include this perspective. An important step in the right direction is Position, Scale and Rotation Invariant (PSRI) object recognition (OR). This form of OR aims to recognise an object independent of its position, scale and orientation. Artificial Neural Networks (ANNs) have, in a short time, made an important impression in this field. Those who have applied ANNs to this subject have loosely followed two different approaches. The first is to remove the variance within the input data in a non-ANN pre-processing stage (Yüceer and Oflazer (2), Grossberg and Bradski(3)), and then present this 'normalised' information to the neural network. The second approach is to take into account the variance in the input data within the architecture of the network. Networks that follow the second approach include higher-order neural networks that incorporate domain specific knowledge within the network architecture (Spirkovska and Reid (4)). The second approach has been limited to supervised learning which is unsuitable for a dynamic process environment where little prior knowledge is available.

We have developed an unsupervised PSRI prototype OR system, IvOR that falls into the first category, but with pre-processing stages with neuro-biological inspirations. This system has proven successful at recognising objects independent of their position, scale or orientation within the viewing domain. An important aspect of our work in the automated recognition of rogue features on cast objects. IvOR 1 has had limited success in differentiating such features on the object. This paper gives a brief description of IvOR 1 and then outlines a modified system, IvOR 2. The approach we have taken in the development of IvOR 2 can be viewed as a cross between the approaches outlined above. It employs structure based on the hypercolumn model of the striate cortex as a feature extractor has been used. A novel technique to elicit relevant features from an image, and then ascertain the correct relationship between the features has been implemented. This technique creates an input to a learning system comprising a modular arrangement of Adaptive Resonance Theory (ART) networks (Carpenter and Grossberg (5)(6) Carpenter et al (7)). A series of ART2a networks categorise and learn object feature representations. Final categorisation is done by an overseeing ART1 network that takes input from the feature representing ART2a modules.

Finally the current state of the work is discussed and the future developments outlined.

INTRODUCTION

The overall aim of our work is a PSRI object recognition system for use in a manufacturing environment, specifically fettling (Balendran et al (8)). Fettling is the process of removing unwanted features (such as runners and risers) from metal castings and forgings. These vary from casting to casting, and are not always present. It is traditionally a manual process carried out in a noisy, polluted environment with the operators subjected to high levels of heat and vibration. These conditions give rise to a number of work related health problems including Vibration-Induced White Finger (VWF) (Taylor and Brammer (9)). Automation is therefore necessary for health as well as productivity reasons. Previous attempts to automate the process have relied on clamping the casting in a jig and passing a robot fettler through a predefined path. This is a far from ideal solution as the robot goes through the sequence whether or not there are unwanted (rogue) features. Our aim is to produce a system that is able to recognise an object irrespective of its position within the work domain, and then compare it with a known standard (i.e. a CAD model or previously presented master object).

The nature of this work lies within the domain of surface matching and 3D feature recognition. Our original approach to the problem was to use the purely mathematical techniques proposed by Besl and Jain (10) and Besl and McKay (11) based on surface matching. These methods proved too time consuming for an on-line manufacturing process. An outcome of this work was the association of one of eight fundamental types of surface primitives (based on Gaussian and mean curvature) with every data point.

This has proved to be a useful procedure as it is invariant of the viewing position, but it is clearly a distinct process from the feature extraction outlined here. A more ideal solution, for purely object recognition problems, would be to work directly from the visual output, typically using a stereo camera arrangement.

IvOR 1

Our initial work in ANN techniques for PSRI OR focused on the development of IvOR 1, an InVariant Object Recognition system that employed a pre-processor based on characteristics of the human visual system and an ART2a NN as the learning/ categorisation system.

Input data

The surface point data that forms the input to the system is generated by a laser line scanner (Keat et al. (12)). Figure 1 shows a representation of a typical data set. A fan laser is directed at the object and viewed from two sides by mirrors and a CCD camera. A Surfa DSP 56xxx laser line scanner is used to convert the line data to point displacement information, which is then converted to X-Y-Z point data based on calibration information. This arrangement is able to collect a maximum of almost 15,000 points per second.

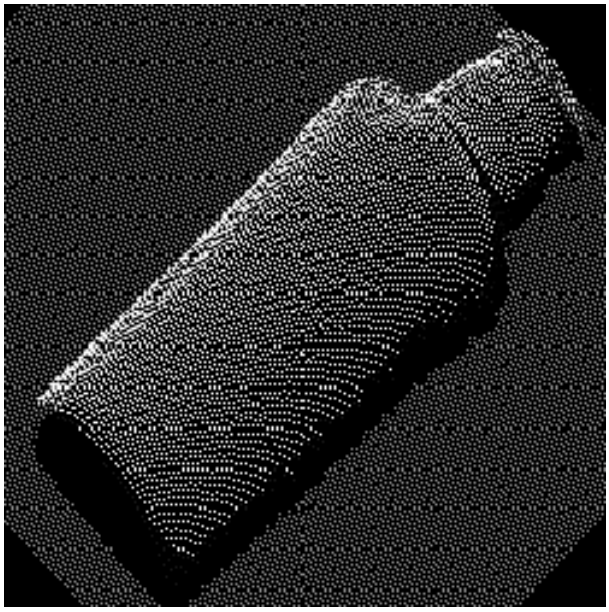

Figure 1 Output from laser line scanner

For every point on the surface (x-y co-ordinate pair), the mean and Gaussian curvatures are calculated using neighbourhood information. The combination of these two curvatures defines the local surface as being one of eight fundamental surface primitives (Besl and Jain (13)). A complete surface type label map is generated for the object based on these defined primitives. Alrashden and Motavalli (14) used Kohonen's Self-Organising Map (SOM) to determine the surface primitives based on the Gaussian and mean curvatures of the local surface areas. This is not an ideal solution as the most time consuming part of the mathematical approach is the calculation of the curvatures. The label map forms the input to the system. Resolution for the labelling can be defined dependent on how much surface detail is required, this also acts as a noise filter.

Retinal pre-processing stage (RPPS)

The first stage in IvOR 1 is a retinal pre-processing stage (RPPS) based on certain characteristics of the foveal region of the human retina. The input elements (receptors), arranged as a hexagonal matrix (Figure 2), receive the input from the surface labelling stage. The size of the matrix is set-up to be larger than the size of the object area. The full retinal matrix has a hexagonal boundary and is segmented into six equal areas. The RPPS determines the number of active elements within each segment. The matrix central position is moved relative to the object until a balance of outputs is achieved for each of the six segments. This process is analogous to the action of the six ocular muscles that move the input image to the centre of the primate retina. At this point the object is centred and the retinal array shrinks to just envelop the object (similar to the process of focusing on an object). The representation is now independent of the objects initial position within the viewing domain.

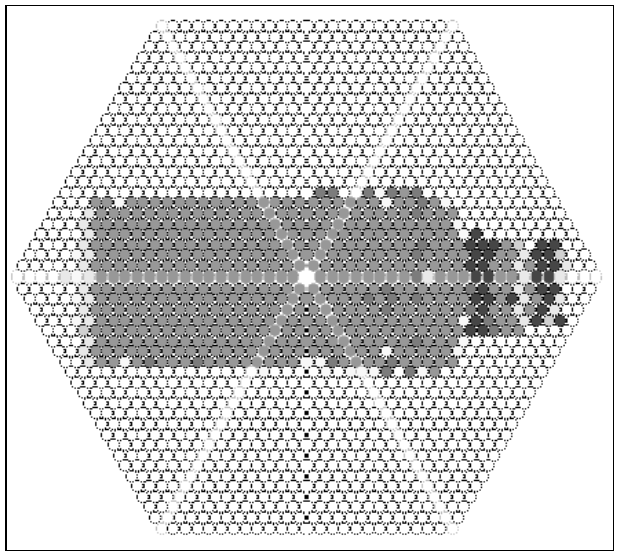

Figure 2 Retinal stage hexagonal array showing different segments and object reduced to surface primitives

The learning and categorisation system (LCS), also arranged as a hexagonal matrix, has a predefined (fixed) number of elements. The number of elements within the RPPS is determined by the size of the input object. The LCS elements have a receptive field whose size is determined by the ratio of matrix size between

the LCS and the RPPS. The result of this stage is that the RPPS always forms the same number of inputs to the LCS and the representation then becomes scale independent.

Learning and categorisation stage

The final stage, the LCS, is based on an ART2a network. The scaled and centred input is fed to the ART2a network. If no match (resonance) occurs on a committed node the input is rotated. This is repeated until a match occurs on a committed node. If no match occurs on a committed node after a full rotation of the input, the input is learnt on an arbitrary uncommitted node.

Rogue feature extraction

The learning mechanism can be deactivated so that the system acts purely as a classifier. In the initial development of IvOR 1, the input was rotated as explained above until resonance (matching) occurred on a committed node of the ART2a F2 layer (i.e. the object was recognised in the orientation that it had been learnt). When this happens the system calculates dummy top-down weights (long term memory (LTM)). These were compared and an output is formed based on the differences between the dummy weights and the LTM for the winning node. Shepard and Metzler (15) have presented experimental evidence to show that there is possible biological plausibility for this method of coping with rotation variance. In their work, subjects were shown identical pairs of 3D object differing only with the angle at which they were observed. They showed a linear relationship between the time required to identify matching pairs and the angular rotational difference between pairs.

The problem with the original IvOR 1 was that resonance would occur on a winning node before complete alignment with the learnt object, even with a high vigilance setting for the ART2a network. This was overcome by finding the point of highest resonance after one complete rotation and then comparing dummy weights and LTM to show differences between the input object and the stored data. The difference between the input object and the memory represents additional (or even missing) features. There is an additional learning and classification system, again using an ART2a NN, for the rogue features. This is not just for the fettling process but is aimed at possible use as a feedback mechanism for the casting process control.

IvOR 1 has proven able to recognise a 3D object independent of its position, scale and orientation within the viewing domain but has limited ability to distinctly differentiate rogue features on the object. This is mainly due to the fact that the rogue features make up only a small part of the input data. Using a complete image requires a large number of inputs to the learning mechanism, we have worked typically with 1800 inputs. We have found ART to be more ‘controllable’ with smaller data sets.

It is hoped to apply some of the principles behind the IvOR 1 system to other recognition tasks where PSRI OR is necessary, the only change necessary would be to define the data pre-processing stage in a format compatible with the RPPS input.

IvOR 2

IvOR 2, the next stage in the development of IvOR, works with object features instead of the complete object. The input stage is identical to IvOR 1, the difference starts at the retinal stage.

Retinal stage

Simulation of the primate retina has been done, in a hardwired form, by Carver Mead and his research group in the development of their silicon retina (Mahowald and Mead (16)). Gaudiano (17) has also looked at software simulation of the interaction between retinal layers. We use synthesised data and therefore a complete retinal simulation is unnecessary. IvOR 2 subjects the input image to an edge detector that has vague parallels with the ‘on-centre off-surround’ receptive field interaction of the bipolar, horizontal and ganglion cells within the retina. Each element in the retinal stage has a receptive field that incorporates all its neighbouring receptors. If this element has a value above or below the average of the elements in its receptive field, plus or minus a given threshold value, the element inputs to the next stage of IvOR 2.

Hypercolumn theory

The retinal stage emphasises regional boundaries within the input. The next stage is the extraction of relevant features. This is done using the simulated arrangement of hypercolumns. The hypercolumn theory for the striate cortex was developed by Hubel and Weisel (18) based on their work mapping the striate cortex of the macaque monkey. The model they produced described the cortex as comprising millions of units known as 'hypercolumns'. Each hypercolumn is made up of six layers and approximately 200,000 cell. Layer 4 (divided into four sub-layers, 4A, 4B, 4C alpha, and 4C beta) accepts input from the retina (via the lateral geniculate nuclei to 4A and 4C). A detailed description of the processing of this information can be found elsewhere (Hubel and Wiesel(18)(19) Hubel (20) Frisby(21)).

Figure 3 shows a single hypercolumn unit as employed

by IvOR 2. The lower section, Feature Extraction (FE), represents the orientation of the feature line. The upper cells, Feature Linker (FL), represent the direction of the end of that line (these cells only fire if the line is an end). A feature is represented on a column when two or more orientation cells' fire. The size of the hypercolumn unit is determined by the desired angular resolution (preferred orientation) for the orientation columns. Hubel and Wiesel (18) predicted a preferred orientation in the macaque monkey of approximately 10 degrees, but this was only because the instrumentation used for measurement of brain activity was unable to detect anything smaller. In our initial work we have used a resolution of 45 degrees to simplify the model.

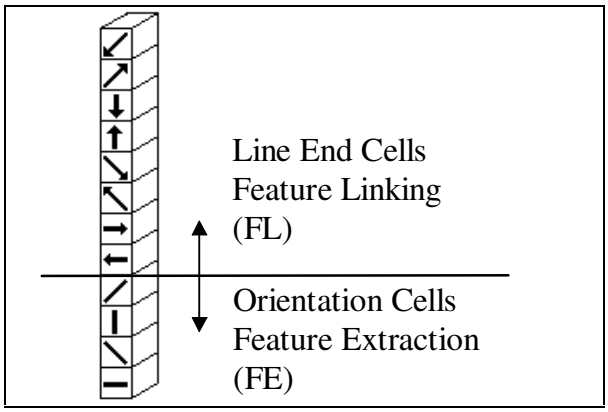

Figure 3 Single hypercolumn unit as used in IvOR 2

Feature extraction using the hypercolumn model

Within the human visual system (including the retina, lateral geniculate nucleus and striate cortex) various parts are sensitive to different stimulations. These include motion, orientation, lines, edges, slits, line ends, colour, and combinations of these. It is obvious from this simple description of the structure of the striate cortex that any attempt to model it would have to be a gross simplification of the biological system. We have, to date, constrained the model to lines, line ends, and orientation. We work with stationary inputs and so motion is not a factor. The initial work has been further constrained to the recognition and categorisation of simple 2D images and surface label maps of simple geometric (3D) objects. Object features have been limited to corners in order to simplify the model. This has proved satisfactory for the simple geometric objects that we have used to test the system. It is our aim to further develop the system for more complex images, and hence features.

Figure 4 shows a simplified representation of the hypercolumn model we have employed. The receptive fields employed are sensitive to lines with orientation resolution of 45°.

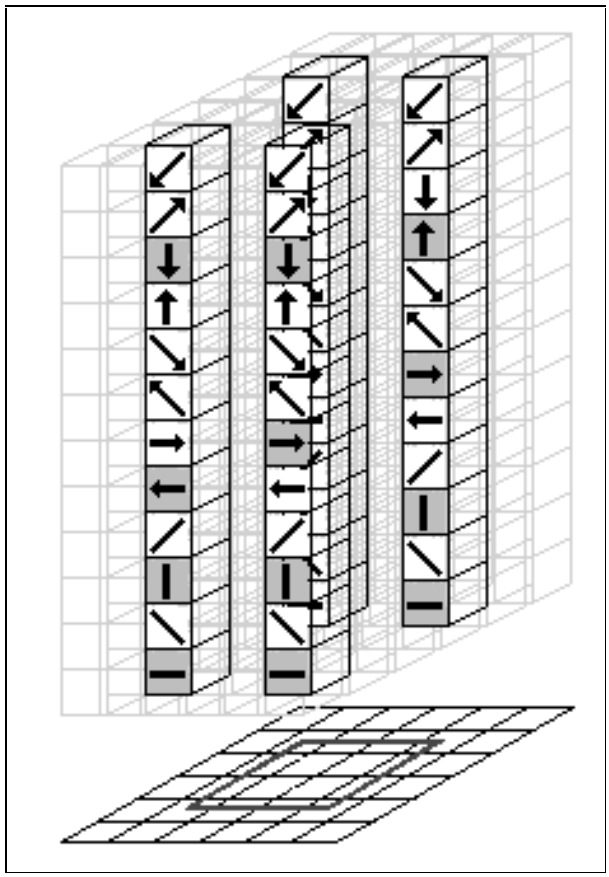

Figure 4 Hypercolumn structure showing firings from simple rectangular object

Feature linking using hypercolumn model

The individual hypercolumns we have used in our model are similar to those adopted by Yamaguchi and Shimizu (22) in their OR system with figure-ground separation. Our model uses a novel method for linking features within the complete hypercolumn structure to give a unique invariant description of an image. If two or more firings are in the FE section, then a significant feature is present at that position (typically two firings represents a corner). Within the model each hypercolumn has connections to the others within the structure through the FL section. These connections are initially all closed. When a feature is detected at a particular hypercolumn a connection is opened to other active hypercolumns. The connection only becomes active when related line end cells within the FL stage of the hypercolumns become connected. This has the effect of ordering the features so that only adjacent features link to each other. Active connections create an input to the learning system based on their relative lengths. Figure 5 gives a representation of this process, with feature point linking. This shows the different interaction between related features.

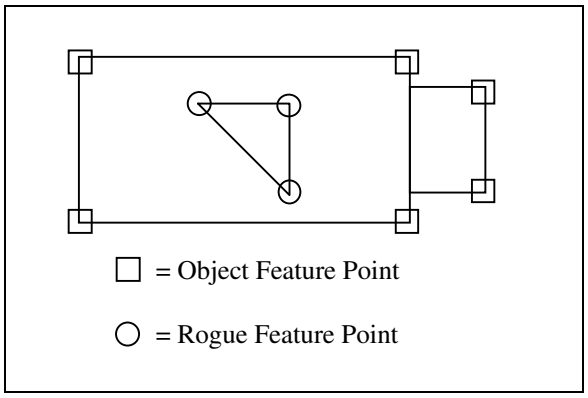

Figure 5 Linking of object and rogue features

The learning and categorisation process

The complete output from the hypercolumn structure is independent of its position and orientation within the input space. The categorisation and learning mechanism employed for the extracted features is a modular arrangement of ART networks (ART1 and ART2a) and is shown in Figure 6. Feature categorisation from the output of the hypercolumn model is done by a series of ART2a networks. Each network deals with different feature types. Our model presented here uses one network to give relative distance of features apart (the scale unit) and a second gives the relative angle between features (the shape unit). The initial layer of the scale unit ART2a network (F0) normalises the input vector, thus creating the required scale invariance for the system. The extent of rotational invariance of the system is dependant on the order in which the feature data is put in the output vector. To create complete rotational invariance the input vector is rotated through all possible angles until a match is found. If no match is found (i.e. no resonance between layers F1 and F2) the ART network learns the inputs for the original presentation. This is similar to the rotational approach taken in IvOR 1.

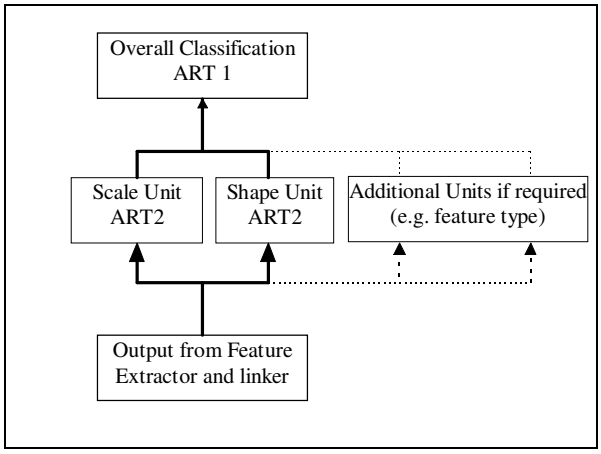

Figure 6 Arrangement of learning and feature extraction modules

When a winning node is found on each ART2a network, the top (F2) layer of each unit network then forms the binary input to an ART1 network that performs the final categorisation.

DISCUSSION

In this paper we have presented two models, IvOR 1 and 2. IvOR 1 has been tested on real life 3D point data and is able to recognise objects in a PSRI manner, but has shown limited success in extracting rogue feature information. The principles behind IvOR 2 have been tested on simple geometric shapes and have been shown to give a representation of the object that is independent of it position, scale and orientation within the viewing domain. It is, at present, suitable only for such objects. The model is too simple to cope with the complexity of objects which IvOR 1 can. To cope with such objects IvOR 2's size, and complexity will have to be increased greatly. The problem of feature extracting and linking, in a figure-ground separation context, has been addressed by researchers in this field (Grossberg and Somers (23), Yamaguchi and Shimizu (22) and others) using coherent oscillations. They have suggested, based on the results of other researchers' physiological experimentation on the cat visual cortex, that there is synchronisation of neural oscillators formed by an object, and unsynchronisation by the background. This enables the linking of features for an object. Looking at these methods may allow a simpler structure for IvOR 2 allowing greater object complexity.

FUTURE WORK

IvOR 2 is still very much at its earliest stages. Future work covers all four stages (input, retinal, hypercolumn and learning) with the aim of allowing the use of input data with a higher complexity, as typified by the fettling problem.

IvOR 1 and 2 have both used surface primitives based on point data as input. This mathematically based technique has been used in many engineering and medical applications. We aim to apply ANN's to the derivation of the primitives based on point data.

The retinal stage has used a simple non-overlapping on-centre off-surround receptive field with only a small field size. We have started to look at using various sizes within the model, and different arrangements of receptors (possibly to model the end-stopping effect utilised by the retina to produce sensitivities to different sized features). The effect of different field sizes and their interaction has been investigated by Jacobs and Kosslyn (24). They showed, through computer simulations, the possibility that if receptive fields are allowed to adapt, they will become smaller for tasks where shape or spatial relations are important, and

larger where they are to encode specific exemplars or metric distances.

Further work will also focus on increasing the complexity of the model to take in a wider variety of features. This will be helped by modular approach taken. The use of individual networks for different feature types allows, to some extent, the possibility of incorporating different network architectures within the complete system. The possibility of using the principle of coherence of neural oscillators (as described by Grossberg and Somers (23) and Yamaguchi and Shimizu (22)) for feature linking will also be considered.

We have more long-term aims for work with the learning mechanism. The representation produced by the hypercolumn model is 'semi-rotationally invariant'. The same representation is produced irrespective of the orientation, but the effectiveness of the categorisation is dependent on the order in which the features are given to the categorisation system. To overcome this problem we cycle through the input vector to the ART network until resonance occurs on a committed node. If no resonance occurs in one complete cycle the input pattern is learnt on an uncommitted node. Our future work includes an adaptation of the F1 layer to take into account rotational invariance in the input to the ART network.

REFERENCES

1. Marr, D, 1982, “Vision”, W. H. Freeman and Company, San Francisco, USA.
2. Yüceer C, and Oflazer K, 1993, "A rotation, scaling and translation invariant pattern classification system", Pattern Recognition, 26, 5, 687-710.
3 . Grossberg S, and Bradski G, 1994, “VIEWNET: A neural architecture for learning to recognize 3-D objects from multiple 2-D views”, SPIE Int Robots and Comp Vis XIII: Alg and Comp Vis, 2353, 266-275.
4. Spirkovska L, and Reid M, 1993 "Coarse-coded higher-order neural networks for PSRI object recognition", IEEE Trans Neural Networks, 4, 2, 276-283.
5. Carpenter G A, and Grossberg S, 1987, "A massively parallel architecture for a self-organizing neural pattern recognition machine", CVGIP, 37, 54-115.
6 . Carpenter G A, and Grossberg S, 1987, "ART 2: Self-organization of stable category recognition codes for analog input patterns", Applied Optics, 26, 3, 4919-4930.
7 . Carpenter G A, Grossberg S, and Rosen D B, 1991, "ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition", Neural Networks, 4, 493-504.
8. Balendran V, Sivayoganathan K, and Howarth M, 1992, "Sensor aided fettling", Proc. 8th NCMR, 1992, 132-136.
9. Taylor W, and Brammer A J, 1982, “Vibration effects on the hand and arm in industry: an introduction and review”, Vibration effects on the hand and arm in industry, John Wiley, New York.
10. Besl P, and Jain R, 1985, "Three dimensional object recognition", Comp. Surveys, 17, 1, 75-145.
11. Besl P, and McKay N D, 1992, “A method for registration of 3-D shapes”, IEEE Trans PAMI, 14, 239-256.
12 . Keat J, Balendran V, Sivayoganathan K, and Sackfield S, 1994, “3-D data collection for object recognition”, Adv in Man Tech VIII, Proc 10th NCMR, 648-652.
13. Besl P, and Jain R, 1988, “Segmentation using variable surface fitting”, IEEE Trans PAMI, 10, 167-192.
14 . Alrashdan A, and Motavalli S, 1994, “Digitized data segmentation using neural networks for reverse engineering”, SPIE Videometrics III, 2350, 299-305.
15. Shepard R N, and Metzler J, 1971, "Mental rotation of three-dimensional objects", Science, 171, 701-703.
16. Mahowald M, and Mead C, 1991, "The silicon retina", Scientific American, May, 40-46.
17 . Gaudiano P, 1993, “Simulation of X and Y retinal ganglion cell behaviour with a non-linear push-pull model of spatiotemporal retinal processing”, Boston Univ Tech Report, CAS/CNS-TR-93-048.
18. Hubel D H,and Weisel T N, 1968, "Receptive fields and functional architectures of monkey striate cortex", J. Physiology, 195, 215-243.
19. Hubel D H, and Weisel T N, 1974, "Sequence regularity and geometry of orientation columns in the monkey striate cortex", J. Comp Neurology, 158, 267-294.
20. Hubel D H, 1988, "Eye, Brain, and Vision", Scientific American Library, New York, USA.
21. Frisby J P, 1980, "Seeing: Illusion, Brain and Mind", Oxford University Press, UK.
22. Yamaguchi Y, and Shimizu H, 1994, "Pattern recognition with figure-ground separation by generation of coherent oscillations", Neural Networks, 7, 1, 49-63.
23. Grossberg S, and Somers D, 1991, “Synchronized oscillations during cooperative feature linking in a cortical model of visual perception”, Neural Networks, 4, 453-466.
24. Jacobs R A, and Kosslyn SM, 1994, “Encoding shape and spatial relations: The role of receptive field size in coordinating complemetary relations”, Cog Science, 18, 361-386.