International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:5, 2010

Artificial Visual Percepts for Image Understanding

Jeewanee Bamunusinghe and Damminda Alahakoon

***Abstract*—** Visual inputs are one of the key sources from which humans perceive the environment and 'understand' what is happening. Artificial systems perceive the visual inputs as digital images. The images need to be processed and analysed. Within the human brain, processing of visual inputs and subsequent development of perception is one of its major functionalities. In this paper we present part of our research project, which aims at the development of an artificial model for visual perception (or 'understanding') based on the human perceptive and cognitive systems. We propose a new model for perception from visual inputs and a way of understaning or interpreting images using the model. We demonstrate the implementation and use of the model with a real image data set.

***Keywords*—I**mage understanding, percept, visual perception.

I. Introduction

THERE has been a large amount of research carried out in Artificial Intelligence (AI) for many decades. A vast number of algorithms and techniques have been developed, which provide 'artificial intelligence' where 'intelligence' is defined in various ways. The techniques that have been most popularly used are neural networks, genetic algorithms, fuzzy logic, bayesian networks, and decision trees.

Intelligence has many definitions and does not only refer to the manipulation of data or information in machines. Issues such as incremental learning, lifelong learning, and concept formation have received the attention of AI researches in the recent past, in their quest for building more 'intelligent' machines. Studying the pinnacle of intelligence 'the human brain' and mind has received a lot of attention also in the recent past to obtain inspiration for AI research. This work has also been called Natural Intelligence (NI) research [1]. Therefore the artificial models which can process visual inputs in similar ways (to human) could contribute immensely to the field of developing useful natural intelligence (or future AI) systems. Within the human brain, processing of visual inputs and subsequent development of perception is one of the major functionalities. In this paper we present a part of our research project, which aims at the development of an artificial model for perception (or 'understanding') based on the human perceptive and cognitive systems.

Jeewanee Bamunusinghe is with Clayton School of Information Technology, Monash University, Australia.
(e-mail: jeewanee.bamunusinghe@infotech.monash.edu.au).
Damminda Alahakoon is with Clayton School of Information Technology, Monash University, Australia.
(e-mail: damminda.alahakoon@infotech.monash.edu.au).

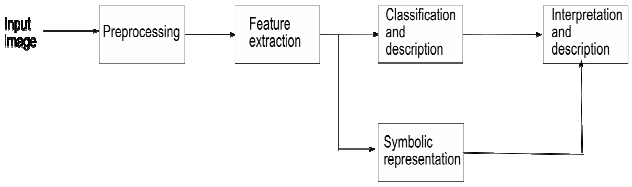

Fig. 1: Components of a computer vision system

Artificial systems perceive the visual inputs as digital images. There have been many algorithms and techniques developed to process and analyse images and other visual inputs over the years. The common approach for image analysis can be summarized as in Fig 1. Clustering is widely used in classification phase of image analysis. Even though the clustering algorithms group the images based on their characteristics, defining and extracting the best clusters require human intervention. And also these algorithms group one image into only one group. With a real image data set a human can categorize one image into many groups as they perform the grouping based on their ' understanding'. In an artificial system maintaining the relationships or links between the groups and the degree of similarity between the groups (inter cluster distances) provide more 'understanding' about the images. Developing a system with 'understanding' can be beneficial for image analysis as it increases 'intelligence' of the machine and reduce the human intervention requirement in image analysis. Rather than the specifics of an image, the proposed model captures general patterns from the presented images, which we present as ‘understanding’ about the environment represented by those images.

In this paper we present an artificial model for developing perception (or 'understanding') from images and explain how this proposed model can be used for image understanding.

In section 2 we present an artificial model for perception from images. How to build percepts from the model is described in section 3. In section 4, architecture for the proposed model is described. The image understanding using

the model is discussed in section 5. Finally section 6 provides a summary of the content of this paper.

II. Artificial Model for Perception

The term perception has been defined in literature [2]- [5]. According to [2]- [4] perception is defined as a process which interprets the sensory information to yield meaningful information. Perception is also viewed as a process by which humans acquire knowledge of an objective world [5].We argue that perception is "a process of interpreting sensory information by using the accumulated knowledge about the world". This interpretation comes from the 'understanding' of the system, not exact pattern recognition.

The proposed model for artificial perception consists of three layers: perception layer, feature layer and dimension layer. Each of these layers are explained in detail next.

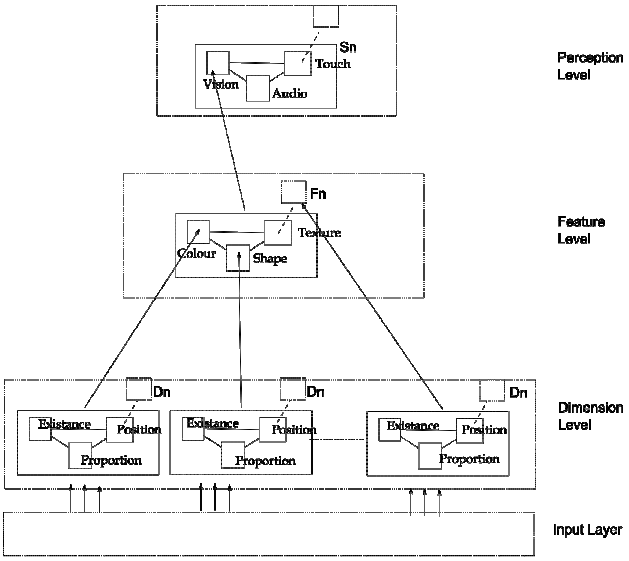

Fig. 2: A model of perception from images

Perception layer: Humans perceive the environment using multiple sources of sensory information derived from several different modalities including vision, audition, touch, olfaction and taste. At any given instance what is perceived is a combination of one or more sensors.

Feature layer: As explained previously each instance of human experience is a combination of multiple sensors. Each sensor input is composed from number of features. As an example visual input can be separated into colour, form, depth and motion. This information is processed in different cortical areas of human brain [6], [7]. Therefore here we assume that these features are independent of each other but are fused together during the actual perception and understanding. Humans have the ability to differentiate events or objects using one or more features.

Dimension layer: Humans differentiate (or identify similarity) objects using a combination of multiple aspects or dimensions. These aspects or dimensions are used (or considered important) according to the environment or context in which the image (or object) is considered. For example, when we use the feature - colour - we can differentiate objects considering the 'existence' of a particular colour in each object. If we had considered the 'proportion' of the existence of colour and the 'position' of existence, a more flexible representation of the images could have been achieved. Since it is not possible to identify a general set of such dimensions to suit every situation, a model should have the flexibility to be tailored for different features.

As described earlier, the dimensions may differ in different situations, and the ability of obtaining this information also needs to be considered. Due to the limitations in current feature extraction techniques we propose three possible ways of explaining images. The model is flexible to accommodate other dimensions.

- By existence of each feature value in image
- By proportion of each feature value in image
- By position of each feature values in image

The links between the nodes in each layer represents the accumulated learning based on past experience. For example, a link could represent how many times red and blue (two colours) existed in images. We consider this model as an abstract representation of images presented to the system, and as such an artificial visual percept representing what is 'seen' by the system.

III. Percept Generation from the Model

In general, the term percept is used to describe an organized sensory experience that carries meaning [4]. The percepts are created by the human brain by extracting the perceptual stimuli from the sensory stimuli. These sensory stimuli are received from multiple sensory modalities such as vision, audition, touch, olfaction and taste. During the perception process, the new percepts can be created or the existing percepts can be updated. When a human perceive a new experience, a new percept is built with available information. These percepts are continuously updated with new inputs. Therefore percepts are considered to be dynamic and can change from moment to moment. In vision, the current visual representation of the dimensions of a scene is called a percept. These dimensions include shapes, sizes, colours, distances and motion.

The definitions of percept derived from the model are given below.

Definition 1 *At perception level, percept (P) is a collection of clusters* CL_{ij} *and their relationship* $R_{ij,kl}$,

where i and k represents the different type of sensor

modality $S_1,S_2....S_n$ and j and l represents the cluster number for sensor S_i : j = 1, 2.....m .

In human perception system these modalities include vision, audition, touch and olfaction.

Definition 2 *At feature level, percept (P) is a collection of clusters CL_{ij} and their relationship $R_{ij,kl}$*

where i and k represents the different type of features $F_1,F_2....F_n$ and j and l represents the cluster number for Feature F_i : j = 1, 2.....m .

For visual perception, these features include colour, motion, depth and form.

Definition 3 At dimension level, percept (P) is a collection of clusters CL_{ij} and their relationship $R_{ij,kl}$,

where i and k represents the different type of dimensions $D_1,D_2....D_n$ and j and l represents the cluster number for dimension D_i : j = 1, 2.....m .

For each feature in definition 2, dimensions include existence, position and proportion.

The steps involved in percept generation, updating and prediction is presented next.

Percept generation: In percept generation a set of images representing the environment is presented. The input can be a single image or a sequence of incrementally evolving images of the same object. However we argue that presentation of sequence of images as incrementally evolving steps gives a complete representation of the object.

Next, from each of these images it is necessary to extract features in all of their dimensions. The available features and the dimensions are limited by the current feature extraction techniques. After the feature extraction, for each feature the feature values are grouped based on their similarity. As an example for colour feature red and green objects, green and blue objects and so on. The number of occurrences of each group gives an idea of frequency of perceiving objects in that group. Similar procedure can be applied to the dimensions of each feature. Let us assume f_x and f_y are two features or dimensions of the image and f_{x1} , f_{x2} ,f_{x3} , f_{x4} , f_{x5} and f_{y1},f_{y2}, f_{y3} are the grouping of images based on features or dimensions.

f_x = { f_{x1} , f_{x2} ,f_{x3} , f_{x4} , f_{x5}}

f_y= {f_{y1},f_{y2}, f_{y3}}

The links between features of the objects are created next. The links can extracted from a co-occurrence matrix obtained from the previous step. The strength of the link represent the number of occurrences of a particular combination. As an example images with f_{x1} and f_{y1} have seen five times.

TABLE I
CO-OCCURRENCE MATRIX FOR FEATURE F_X AND F_Y

	f_{x1}	f_{x2}	f_{x3}	f_{x4}	f_{x5}
f_{y1}	5	0	1	3	1
f_{y2}	0	0	2	5	1
f_{y3}	6	0	1	4	1

Prediction: When a new image is presented to the system, the system should provide an opinion about the new image (interpretation). In this section we discuss the how the model carry out prediction.

During the prediction process, as in the percept generation phase the features and dimension values are extracted from the new image. It is then necessary to decide the level of detail for prediction. If it is required to give an opinion in feature level, feature values are used and to give an opinion in dimension level dimension values of each feature are used. Let us assume new feature or dimension values are x and y for f_x and f_y respectively. For each feature or dimension the model calculates the closest existing value from f_x and f_y which are f_{xi} and f_{yi} respectively.

```
if (x== fxi) and (y==fyi) then
   "I have seen this before"
else if (x== fxi) and (y!=fyi) then
   "I have seen an object with x, but fy is fyi"
else if (x!= fxi) and (y==fyi) then
   "I have seen an object with y , but fx is fxi"
else
   "I haven't seen this before"
```

Not only that, the system can also provide an opinion about images that are close to the input image in several other dimensions or features.

Updation: Updating the existing knowledge should happen every time the system perceives a new input so that the system can enhance its understanding about the environment. During this process new feature values, dimension values and the links can be added. As a result the co-occurrence matrix generated in the percept generation phase is updated. For this phase the information extracted from the image in the previous section can be used.

```
if both x and y exists in fx and fy then
   create or update the link between x and y
else if x exists and y do not exists
   add y into fy and create a link between x and y
else if y exists and x do not exists
   add x into fx and create a link between x and y
else if both x and y do not exists
   add both x and y into fx and fy and create a link between
x and y
```

IV. Architecture of the Model

The model presented in section 2 can be implemented using existing techniques in image processing and AI. In this architecture, the sequence of steps follows the same order as the human visual system. The proposed architecture is presented in Fig. 3. Here the environment is represented using a set of images. In humans this comprises of visual fields. The images are then pre-processed using image processing techniques to enhance the image quality for feature extraction. Subsequently features are extracted from images. This step represents processing at subcortical areas and the early visual pathway of humans to a certain extent. In the human vision system, the cortical processing begins next. On the basis of anatomical and physiological studies, many regions of human cortex are involved primarily or extensively in the processing of visual information. The Self Organizing Maps (SOM) proposed by T. Kohonen [8] are widely used to represent the computational maps of human brain in computers. This algorithm captures the essential features of brain information processing. SOM is widely is used as a clustering technique. Thus it is assumed that the SOM based clustering can be used to artificially simulate groupings formed in human brains. In this architecture SOM based clustering algorithm is used next to generate groupings of images based on required features and their dimensions. Finally the percepts are built using the information available from the previous step. In the human vision system this step is assumed occur as a psychological process. The percept includes the clusters, the links between the clusters and frequency of grouping and links.

A. MPEG-7 Descriptors

The moving picture expert group (MPEG) was established in 1988 to develop standards for coded representation of moving pictures, associated audio and their combinations. The first set of standards was released in 1992 with a nickname of MPEG-1. Later several other versions of these standards were issued with the increased quality (MPEG-2, MPEG -4). In this paper MPEG-7 descriptors were used for feature extraction. Widely used MPEG-7 descriptors are shown in Fig. 4.

A. Growing Self Organizing Map

Clustering is the unsupervised classification of patterns (observations, data items or feature vectors) into groups (clusters) [9].The similarity between two patterns drawn from the same feature space is measured using a similarity measure. The most popular metric for similarity measure is *Euclidean distance*. The Euclidean distance in an m-dimensional feature space:

$$d_2(x_i, x_j) = \left(\sum_{i=i}^{m} (x_{ik} - x_{jk})^2 \right)^{1/2} \quad (1)$$

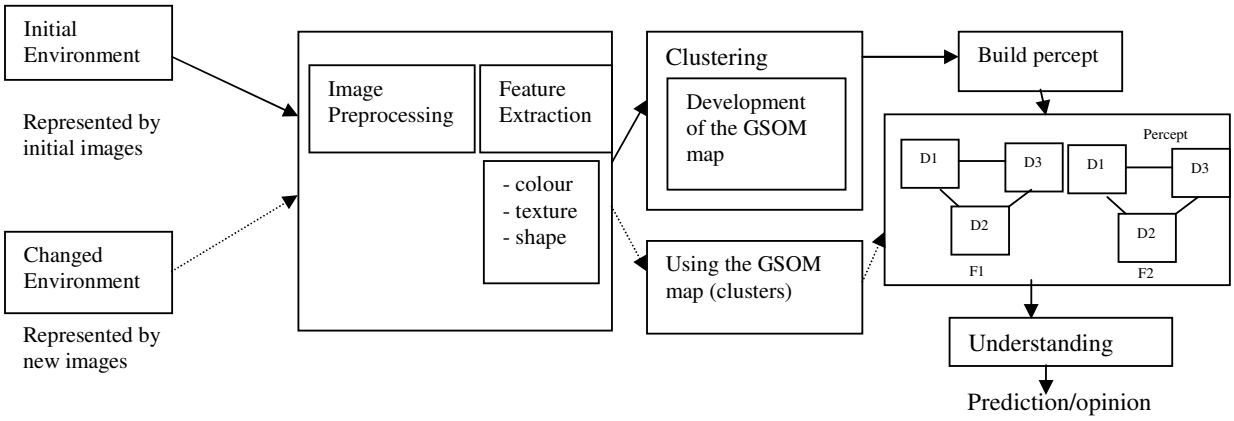

Fig. 3 Architecture of the model

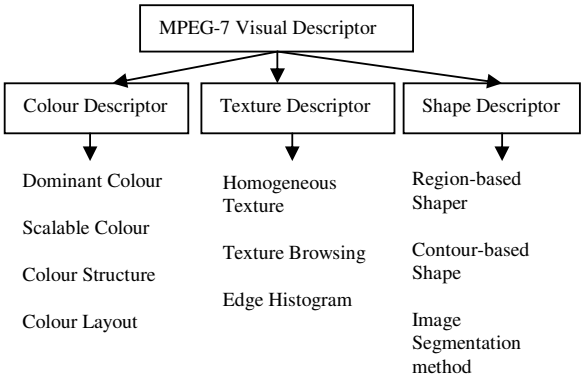

Fig. 4: MPEG-7 Visual Descriptors for low-level features

Competitive neural networks (or winner-take-all) have been used extensively over the past few decades for data clustering. In Competitive learning, similar patterns are grouped by the network and represented by a single neuron. In this paper Growing Self Organizing Map [10] was used as the clustering tool which is an extended version of SOM, a well-known competitive learning algorithm for data clustering which was briefly describe earlier.

The Growing Self Organizing Map (GSOM) is an unsupervised structure adapting neural network algorithm for data mining. This algorithm follows the basic concepts of SOM but has a dynamic structure. The initial map starts with minimal number of nodes (usually four) and grow if required. Unlike the SOM, the GSOM develops into different shapes depending on the clusters present in data. Thus the GSOM provides a better representation of the data set. The two dimensional grid structure of the GSOM is maintained by adding new nodes only at the boundary nodes of the network.

The growth of the GSOM is controlled by the parameter called Spread Factor (SF). Using the SF, the data analyst has the ability to control the growth of the GSOM. A lower spread factor produces an abstract overview of the input data by creating smaller maps while the higher spread factor gives a detailed analysis of the data using larger maps. The spread factor is independent of the dimensionality of the data set.

V. Image Understanding Using the Model

Image analysis is concerned with the extraction of measurements, data or information from an image by automatic or semiautomatic methods. This field is has been called by a variety of other names including scene analysis, image description and image understanding. Image analysis is not limited to the classification of scene regions into a fixed number of categories, but rather are designed to provide description of complex scenes.

Clustering has been widely used in the classification phase of image analysis. Even though the clustering algorithms group the images based on their characteristics, defining and extracting the best clusters require human intervention. And also these algorithms group one image into only one group. In a real image data set humans can categorize one image into many groups as they perform the grouping based on their ' understanding' as well as their interests and background. As an example the same image data set can be grouped by different features (colour, texture...) and by different dimensions (existence of the feature, proportion, position). Also humans have the ability to select appropriate features and dimensions for groupings based on a specific situation. Existing image processing systems lack this capability. The proposed model is an attempt to fill this gap.

In this section we demonstrate how our proposed model can be used for image analyse using a real data set. In this study a dataset originally created by University of Washington for the purpose of content-based image retrieval was used. It contains the colour images of sceneries such as beaches, buildings and people, forests, snow and sea. In this image database, each image is labeled with their content. e.g: trees-bushes-grass, beach-ocean-sky-cliffs. In this experiment 60 images were selected from the original dataset.

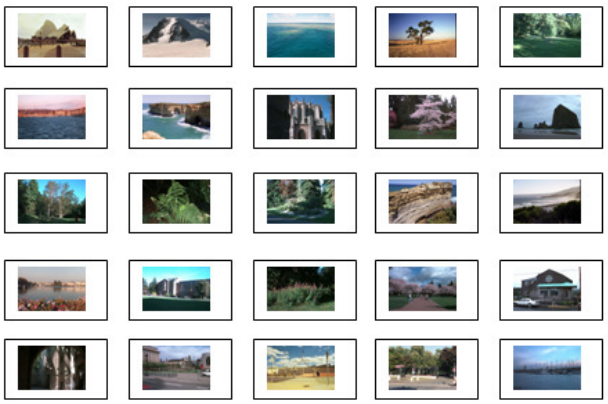

Fig. 5 A data set of sceneries extracted from a collection provided by University of Washington [11].

The images can be analysed using a number of different features of the images. Colour, texture and shape to name a few. In this experiment we reduce our scope to only the colour and texture features.

Feature extraction: In this experiment, two MPEG-7 visual descriptors were used to extract the colour and texture features from the images. The feature extraction was performed using Visual Descriptor Extraction (VDE) toolbox developed by Image and Video Analysis (IVA) research group and Caliph Toolbox. A brief overview of each descriptor is presented below.

- Colour Structure Descriptor: The CSD captures the localized colour distribution of an image using a small structuring window. It counts the number of times a particular colour is contained within the structuring element as the structuring element scans the image.

- Dominant Colour Descriptor: The DCD captures the representative colours present in a region of interest in an image. This feature descriptor provides dominant colours, their percentage values, special coherency of the dominant colour, and colour variances.
- Edge Histogram Descriptor: The EHD captures the spatial distribution of edges in an image. EHD subdivide the image into 16 sub images and edges in each block is categorized into one of the following five types: vertical, horizontal, 45^0 diagonal, 135^0 diagonal and non-directional.

According to the proposed model for perception, images can be interpreted or described in three dimensions; by existence, by proportion and by position. Thus in this experiment we extracted different colours and textures present in each image, their percentages and the positions. Using the colour structure histogram we extracted the colours present in each image and their percentages. We used a threshold (more than 50 counts) to extract the major colour components of the image. In order to extract the colour layout of the image we chopped the image into 64 equal blocks and the dominant colour of the each block was obtained. The edge histogram descriptor was used to extract the texture information from images.

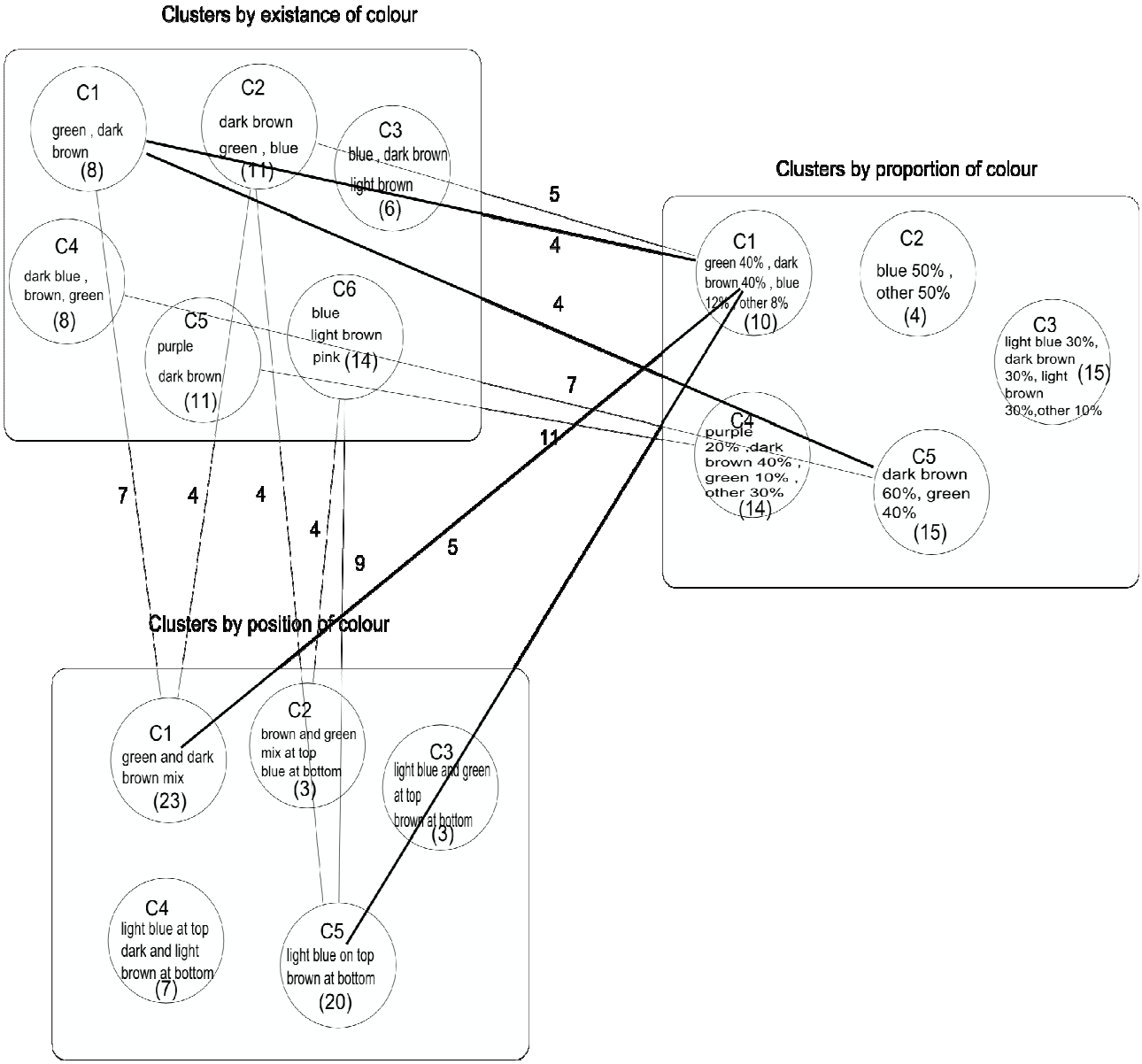

Fig. 6 Percept at dimension level for colour dimension

TABLE II
THE PROCESS OF IMAGE UNDERSTANDING FROM THE MODEL

Component	Functions	Result	Description
Feature(colour), Dimension(Existence) , cluster (C1)	Generate List1	List 1 • bushes-flowers • bushes-flowers • bushes-flowers-trees-grass • clear-sky-tress-bushes-grass • clear-sky-tress-bushes-grass-people - dogs • tress-bushes-flowers-grass • tress-flowers • tress-flowers-grass	Existence of colours (green , dark brown)
Dimension (Existence), cluster (C1) Dimension (Proportion), cluster (?)	Identify related clusters to Dimension(Existence) , cluster(C1) in Dimension(Proportion)	Dimension(Proportion) ,cluster(C1) and Dimension(Proportion) , cluster(C5)	
Dimension(Proportion) , cluster(C1) and Dimension(Proportion) , cluster(C5)	Generate list 2 and list 3	List 2 • clear-sky-tress-bushes-grass • clear-sky-tress-bushes-grass-people-dogs • tress-flowers • tress-flowers-grass List 3 • bushes-flowers • bushes-flowers • bushes-flowers-trees-grass • tress-bushes-flowers-grass	Green 40% , dark brown 60% Green 40% , dark brown 40% , blue 12%, other 8%
Dimension(Proportion) , cluster(C1) and Dimension(Proportion) , cluster(C5)	Difference between Dimension(Proportion), cluster(C1) and Dimension(Proportion), cluster(C5)	Blue and other	Blue exists in list 3 in addition to green and dark brown
Dimension(proportion) , cluster(C1) Dimension(Position), cluster(?)	Identify related clusters to Dimension(proportion) , cluster(C1) in Dimension(Position)	Dimension(position), cluster(C1), cluster (C2) and cluster(C3)	
Dimension(position), cluster(C1), cluster (C2) and cluster(C3)	Generate list 4 , list 5, list 6	List 4 • bushes- flowers –rocks-grass • clear-sky-tress-bushes-grass • trees-bushes • trees-flowers • trees-flowers-grass List 5 • clear-sky-tress-bushes-grass • clear-sky-tress-bushes-grass-people-dogs List 6 • Partly-cloudy-sky-building-tree-tree-trunk-bush-people-grass-HUB-area • Partly-cloudy-sky-tree-leafless-tree-grass-plant	Green and dark brown mix Light blue and green at top and brown at bottom Light blue on top , brown at bottom

Percept generation: The output from the previous stage (6 vectors for each image representing the existence, proportion and position of colours and texture) is then clustered separately using the growing self organizing map. Fig. 6 shows the clusters generated by the GSOM for colour feature and some key patterns that were stored in the artificial percept model after the input of the 60 images. By analysing the information we can see that certain patterns from the images have been 'self learnt' by the system. The groups by dimension (existence) shows the different colour combinations and similarly the other groupings show the partitioning by proportions and the positions of colours. The weights for the links were obtained using co-occurrence matrix between clusters. The numbers in each cluster represents number of images grouped in to that cluster. A similar diagram can be obtained for the texture feature (feature (texture)).

Fig 7: Images belong to cluster 1 of colour

Interpretation and understanding: According the architecture proposed in the Fig. 3, our model is capable of capturing general patterns in the images and providing interpretations about the images "seen" (presented to) by the system. Table 2 demonstrates results of interpreting selected sections of Fig. 6. Based on Fig. 6 the system can identify that there is a group of (8) images which are mainly of colours green and dark brown. By using the proportion dimension, it can also identify the above group an b separated into 2 main groups. One with green and dark brown and the other with an additional blue (12%) colour. Using the position dimension it could be identified that the blue colour exists in the top part of the images. As such the model can capture and store general patterns from the images. As another example, from the position dimension, it could be seen that all blue colour exist in the top part of the images. This is due to the fact that in the 60 images used, the ones with blue colour were those with sky or sea in the top part of the images. The model has the flexibility of analyzing the images presented with one or more dimensions, which can be compared to the human ability of identifying patterns from different dimensions.

VI. Conclusion

This paper describes a way of analysing images using an artificial model of human perception compared to traditional image analysis. The use of the model for image processing and analysis was demonstrated using a real data set consisting of 60 images. The model provides the structure for accumulating information from the images presented to the system and provide a self learnt description of the images. This experiment can be further extended with multiple features and more dimensions to obtain enhanced analysis of images. We are currently working with other image data sets to identify such general patterns and also to compare these results with human pattern identification ability.

References

[1] S. M. Potter, What can Artificial Intelligence get from Neuroscience? Springer-Verlag, 2007, pp. 174–185.
[2] K. M. Galotti, Cognitive psychology : in and out of the laboratory , Thomson/Wadsworth, 2008.
[3] P. O. Haikonen, The Cognitive Approach to Conscious Machines, Imprint Academic, 2003.
[4] M.B. Howes, The psychology of human cognition, Pergamon Press, 1990.
[5] B. Maund, Perception, Central problems of philosophy, Acumen Publishing Ltd, Chesham, [Eng.], 2003.
[6] E. R. Kandel, J. H. Schwartz, and T.M. Jessell, Principles of neural science, McGraw Hill, 2000.
[7] M.F. Bear, B.W. Connors, and M.A. Paradiso, Neuroscience: exploring the brain, Philadelphia : Lippincott Williams and Wilkins, 2007.
[8] T. Kohonen, Self-organizing maps, Berlin, New York: Springer, 2001.
[9] A.K. Jain, M.N. Murty and P.J. Flynn, 'Data Clustering', ACM Computing Surveys, 31(3) , 264-323,1999
[10] D. Alahakoon, S.K. Halgamuge, and B. Sirinivasan, 'Dynamic Self-Organizing Maps with Controlled Growth for Knowledge Discovery', IEEE Transactions on Neural Networks, 11(3), 2000, pp. 601–614.
[11] University of Washington, Content-based image retrieval database. Website, http://www.cs.washington.edu/research/imagedatabase/groundtruth/.