World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:4, 2008

Enhancement of Stereo Video Pairs Using SDNs To Aid In 3D Reconstruction

Lewis E. Hibell, Honghai Liu and David J. Brown
Institute of Industrial Research
University of Portsmouth
Portsmouth, PO13QL
UK
lewis.hibell:honghai.liu:david.j.brown@port.ac.uk

Abstract**—** This paper presents the results of enhancing images from a left and right stereo pair in order to increase the resolution of a 3D representation of a scene generated from that same pair. A new neural network structure known as a Self Delaying Dynamic Network (SDN) has been used to perform the enhancement. The advantage of SDNs over existing techniques such as bicubic interpolation is their ability to cope with motion and noise effects.

SDNs are used to generate two high resolution images, one based on frames taken from the left view of the subject, and one based on the frames from the right. This new high resolution stereo pair is then processed by a disparity map generator. The disparity map generated is compared to two other disparity maps generated from the same scene. The first is a map generated from an original high resolution stereo pair and the second is a map generated using a stereo pair which has been enhanced using bicubic interpolation. The maps generated using the SDN enhanced pairs match more closely the target maps. The addition of extra noise into the input images is less problematic for the SDN system which is still able to out perform bicubic interpolation.

Keywords**—** Genetic Evolution, Image Enhancement, Neuron Networks, Stereo Vision

I. Introduction

THE frontal placement of the eyes in human beings and other animals allows the use of stereo vision. This arrangement of eyes is found in hunter animals as it gives increased accuracy, albeit at the expense of a large field of view (FOV) [2]. Other animals which have lateral placement of the eyes experience a large FOV which is essential for those which are themselves hunted as the approach of a hunter can more easily be detected [3]. However, this large FOV cannot be manipulated like a frontal eye placement view to generate a detailed view of a scene. One form of manipulation performed using the frontal placement of the eyes is fusion in order to generate a 3D representation of the scene. It is this fusion that equips animals with frontal eye placement with a high level of depth perception. The differences between the left and right view of an object are used to calculate the distance of that object from the observer and other objects in the scene. This can be thought of as a form of subconscious triangulation based on the fixed distance between the eyes and their relative angles [4].

Stereo vision is an efficient method of generating a 3D interpretation of the world, however there are many other depth cues which can be used to judge distance in a single image such as shadowing and parallax effects. The representation of the world human beings visualise if based on a combination of all these different depth cues [5], [6]. However it is possible for the human mind to ignore some of the individual cues if they are not consistent with the norm. An example is the use of anaglyphs. Stereo images captured of a scene can be presented to an observer so that each eye sees the correct data and the 3D scene can be observed. The most commonly recognisable form of anaglyphs being the "Red/Blue" variety which became popular in theatres. The stereo information alone is enough to allow the brain to recreate a 3D representation of the scene. Other cues such as the relative angle of the eye and focal length are overridden to allow the 3D scene to be viewed. This shows the power of the information present in stereo views over other depth cues.

The majority of stereo vision related applications using computational artificial attempt to combine views from stereo cameras to recreate 3D representations of the scenes. The small differences between the scenes captured with the left and right cameras can be used to generate a disparity map. The generation of this disparity map can be non-trivial due to effects such as occlusion in the scenes. These 3D representations can either then be used to generate a 3D model of the scene or may be used to calculate specific information such as object ranges. If the stereo views are of a higher quality before attempting the 3D reconstruction the algorithm which generates the reconstruction can obtain higher quality results.

SDNs work by using multiple low resolution frames of a subject in order to produce a high resolution single frame, [1] The combining of multiple frames from a video source in order to generate a higher resolution single frame was pioneered by Tsai and Huang in [7]. In this work they outlined a method for aligning and then combining multiple satellite images in order to generate higher resolution scenes . They outlined a frequency domain approach, this attempts to combine the frequency content from the different low resolution frames. The other main class of approach is the spatial approach, which attempts to perform the enhancement without transforming the

The authors would like to thank the EPSRC for their financial support under grant No. GR/P04220/01. The authors would also like to acknowledge the help and support of Ray Stead from Portsmouth City Councils CCTV for his invaluable help during this research. Earlier work relating to this research can be found in [1]

images into frequencies.

Other methods for multi-frame enhancement expanding on the work by Tsai and Huang include [8]–[11] who use Bayesian, back projection and preconditioned conjugate approaches respectively. A useful paper when looking for what has been done so far in the field of super-resolution is [12]. This gives a description of different types of super-resolution techniques and also outlines those aspects which have not received sufficient attention. However, it does date from 1998 making it more useful as an indication of the field at the time than as a complete modern view.

The use of Recurrent Neural Networks (RNNs) to perform multi-frame enhancement was investigated by Salari and Zhang in [13]. Miravet and Rodriguez [14], [15] also use neural networks for multiframe enhancement.

Many papers regarding multi-frame enhancement are based on the two main underlying methods, POCS (Projection Onto Convex Sets) based and MAP (Maximum A-Posteriori). Borman and Stevenson offer detailed analysis of both methods in their review paper [12]. This review paper also explained the initial developments which occurred after the seminal Tsai and Huang paper, assessing both the advantages and disadvantages of Frequency and Spatial domain approaches as well as other techniques.

The following section describes the networks used to perform the enhancement. Section III explains how the SDNs are used to improve results from a disparity map generator, section IV shows the results of this application in comparison with a traditional enhancement method. Section V gives some concluding remarks regarding the results and offers some possible extensions for future work.

II. SDNs

Self Delayed Dynamic Networks SDNs were created with the intension of processing temporally sampled data with correspondence between time samples. SDNs are forward flowing network structures with the ability to store inputs at time t and use them at time $t+n$ (where n is a number of time periods later). The value of n is learned during training and can vary if necessary for each input. This allows the useful values from different time periods to be accumulated as they appear and then combined to calculate a more accurate result. They were designed to allow pixels from different temporal frames from a video sequence to be intelligently combined to form a high resolution image. Their structure is similar to traditional neural networks with an input layer, an output layer and several hidden layers. However, the algorithm used to propagate the network inputs towards the network outputs has been created to allow the combination or selection of pixel values. They feature several different types of parameters which control the flow of data on the links, however none of these parameters can change the magnitude of an input value. The outputs of the networks are produced using the timing of the inputs combined with their relationships to each other. The network's dependence on its temporal structure and lack of feedback elements make it unsuitable for training with gradient descent or similar methods. Training of these parameters is performed using a modified genetic algorithm (GA) detailed later based on traditional GA techniques. The main elements of SDNs which differentiate them from existing neural networks are outlined below.

Transfer Functions: Transfer functions within neural networks are usually chosen both for their ability to produce a smooth output and their ability to be easily differentiated during application of training methods. Many of these functions also have limited output ranges (0 to 1 or -1 to 1), this stops the values within the network becoming too large and causing it to become saturated. However, when dealing with pixel values it is important to ensure their consistency as they traverse the network. If attempting to enhance a set of inputs which are a specific value the outputs should also consist of those values. Therefore the transfer function used within the nodes of an SDN is a straightforward averaging of the input values i.e. $y = \frac{\sum_{i=1}^{I} x_i}{I}$ where y is the node output, x_i is the i^{th} node input of the I total inputs. This ensures the output of a node is within the range of the input values given. The Gate Control Parameters (GCPs) and thresholds within the network control the x_i values to avoid unnecessary averaging. If SDNs were to be used in the future for another task which does not have such a reliance on the preservation of input values the transfer function could be changed to suit.

Weights: Traditional neural networks feature weights on each link between two nodes. This weight is used to scale the value passing along the link before it is subject to the transfer function of the node $F()$. This leads to the use of Eq. 1 yielding the output of the node, where Y is the node output, X_n is a node input and W_n is weight value taking the form of any real number.

$$Y = F(\sum_{n=1}^{N} X_n W_n) \tag{1}$$

As stated above one of the goals of the SDN approach is to preserve input values wherever possible. Multiplicative weights have therefore not been used in SDNs. The role of weights within SDNs is performed using the Gate Control Parameters (GCPs) and thresholds. These GCPs and thresholds are used to dictate which values are passed to the node. This results in a function which is mathematically similar to that in Eq. 1 but with the real valued weight matrix being replaced with a matrix containing only 0s or 1s. The result of which is to use only selected weights in the sum. Thresholds are also used within the network to supplement the GCPs and allow inputs to be routed or discarded appropriately.

Biases: The necessity of the SDN to preserve input values also means that traditional bias values are not implemented. Bias values are used within standard neural networks to act as an extra input to a node. This extra input is used to adjust the total summed input to the network as shown in Eq. 2 where

B is a bias value and other parameters are as above.

$$Y = F(\sum_{n=1}^{N} X_n W_n + B) \tag{2}$$

Any bias value that was used within an SDN would cause a shift in the node input which was not directly related to pixel intensity. This ability may be useful in other imaging applications (such as contrast adjustment) but is undesirable if the SDN is attempting to preserve the intensity values.

Output Generation

Outputs are produced from an SDN based on the paths input pixels take when traversing the nodes and links. These paths are controlled by the gate control parameters and thresholds (S, C and T) and the iteration number t. Generating the blocking matrix B_{nti} allows the solutions to many conditionals to be condensed into a single matrix.

$$B_{nti} = \begin{cases} 0 & \text{if } S_n < t < 2S_n \\ 0 & \text{if } C_{ni} < t < 4C_{ni} \\ 1 & \text{if } |y_{ti} - \text{avg}(I_t)| < T_{ni} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Where B_{nti} is a blocking matrix value for an input i to node n, S_n is the node's self gate control parameter, C_{ni} and T_{ni} are the connection control parameter and threshold between nodes n and i. y_{ti} is an input from node i, I_t is a subset of inputs coming from the previous layer which have satisfied the S_n and C_{ni} conditionals.

$$\begin{bmatrix} y_{t1} \\ . \\ . \\ . \\ y_{tN} \end{bmatrix} \times \begin{bmatrix} B_{nt1} \\ . \\ . \\ . \\ B_{ntN} \end{bmatrix} = \begin{bmatrix} X_{nt1} \\ . \\ . \\ . \\ X_{ntN} \end{bmatrix} \tag{4}$$

Where N is the number of nodes in the previous layer, y_{ti} is an output from node i in the previous layer, B_{nti} is the blocking matrix for node n and X_{nti} is the value based on y_{ti} which will arrive at node n. The outputs from the previous layer y_t are combined with the blocking matrix for a specific node B_{nt} to create the inputs to the node X_{nt} which may be written as a stacked set of equations, $y_t \times B_{nt} = X_{nt}$.

$$Y_{nt} = \begin{cases} Y_{n(t-1)} & \text{if } \sum_{i=1}^{N} X_{nti} = 0 \\ f(X_{nt}) & \text{otherwise} \end{cases} \tag{5}$$

Where Y_{nt} is the output of node n, N is the number of nodes in the previous layer, f is the node function, X_{nti} is the input from node i to node n. The output of the node Y_{nt} is either recalculated or preserved. The network performs these steps a number of times, denoted here by t.

Self GCP S: The Self GCP S is used to ascertain when recalculation of a node output should be undertaken. If the current iteration t is between the node GCP and double the node GCP then a recalculation is performed otherwise the output remains unchanged.

Connection GCP C_i: The Connection GCP C_i is used to determine if a value coming from node i is ready to be used in the calculation of a node output. If the iteration t is between the connection GCP and four times the connection GCP then the value coming from i is included in the calculation otherwise it is ignored.

Threshold T_i: This is used to judge if the value coming from node i is within a certain range of the other values being used in the node.

Training

Due to the nature of SDNs and their timing based lifecycle structure a Genetic Algorithm (GA) was created for training. This evolutionary approach also allows the combination and further evolution of networks trained on different images by creating a population containing both solutions. This mixed population can create new networks featuring the best elements of each solution.

The chromosomes within the genetic algorithm population contain the parameter values used in the construction of an SDN. The chromosomes are not strings of bits as used in traditional Genetic Algorithms [16] but instead are a 2D matrices of integers. Each integer value represents a parameter value in the network.

III. Stereo Enhancement

In its context here stereo refers to the use of two image capture devices used in a paired manner in order to later recreate a scene in three dimensions. This is possible because the two views of the scene captured from slightly different locations at the same time. Algorithms are available [17] which can generate from the differences in these views a partial 3D reconstruction of the scene. The interest in this technology here is the capability of enhancing the 3D scene which is generated. Stereo reconstruction is just one of many other image processing techniques which can benefit from the use of pre-enhancement. The benefit of using images enhanced using SDNs as the inputs to other processing techniques are the accuracy of the enhancements. If the images processed by an algorithm are blurred it can be more complicated for them to process the information. The results from the SDN enhancement are less blurred and also contain less noise which allows any further techniques to function in a more accurate manner.

The same network which can enhance members of a frame sequence can also be utilised to enhance 3D views originating from a stereo camera pair. If the left and right sequences are enhanced by treating each as an individual frame sequence high resolution versions of both left and right views can be generated. Although the focus here is on the utilisation of two frames the theory is expandable for use alongside any multi-camera 3D scene generation application. These high resolution left and right views can then be used to generate a high resolution 3D view as shown in Fig. 1. The advantages of enhancing the views at an early stage is that the 3D generation algorithm will be able to produce a more accurate 3D representation.

Future work involving 3D enhancement could utilise an SDN which accepts voxels (Volumetric Pixels) and routes them to correct locations in a resultant 3D representation, this is mentioned briefly in section V.

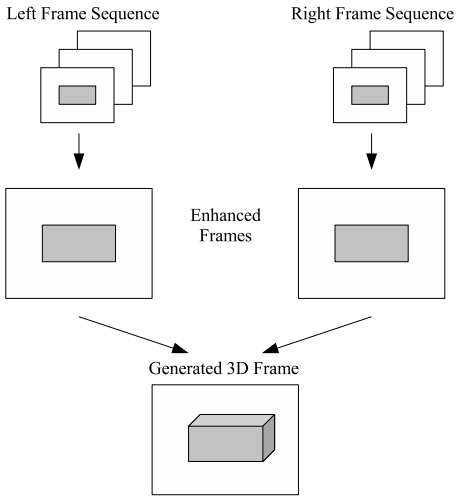


Fig. 1. Stereo Scene Generation

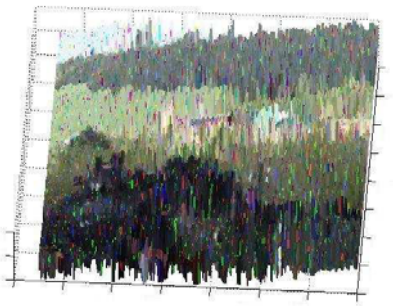
Why Enhance Stereo

Enhancement of stereo images is advantageous if you wish to use the generated depth information for further analysis tasks. This can allow more robust object segmentation within a scene such as at [18] or could be used to allow geologists to observe changes in a landscape over time. After segmentation has occurred a system could take the stereo source images as input and generate a higher resolution view of the scene which can then be used to provide high resolution views of the identified objects.

IV. Results

Results in this section relate to the enhancement of scenes represented in three dimensions. The 3D scene representations to be enhanced are generated here using two views of the scene captured using stereo cameras. The disparity which exists between these two views of the scene are used to estimate the distances within the scene required to reconstruct a partial 3D reconstruction of the scene. The enhancement was performed on the two scenes separately before they are used for 3D scene generation. This allowed the 3D scene generation algorithm to process a higher quality pair of images and therefore generate a higher quality 3D representation. The results in Fig. 2 were generated by firstly enhancing the left and right views separately. These enhanced left and right views where then passed to a disparity map generator and an occlusion estimator [17]. This disparity and occlusion information is then used to generate a depth-map relating to the 3D scene. This depth-map indicates the distance of a pixel from the left camera in relation to all other pixels. The left view was then overlaid onto this depth-map to allow the scene to be viewed as a 3D image. The stereo images generated using this method can be

(a) SDN Result



(b) Bicubic Result

Fig. 2. 3D Image Reconstruction

seen in Fig. 2(a) and Fig. 2(b). Fig. 2(a) was generated using the SDN enhancement of the left and right views and is more accurate than that generated using bicubic interpolations of the left and right views shown in Fig. 2(b).

a) Scoring: Although scores could be given for the accuracy of the enhanced left and right views compared with each of their respective targets this would not show the benefit of using SDN before a 3D reconstruction step. Scoring is therefore based on the differences between the depth-map generated using the original left and right views and a depth-map generated in the following manner. Three left views are down graded to produce one set of inputs for the SDN. This is then enhanced in an attempt to recreate the original left view of the scene. This is put to one side and the same process repeated with three views from the right. The enhanced left and right view are then used to generate a depth-map. The depth-map from the enhanced view is then compared to the depth-map from the original views of the scene. A depth-map is also generated from the bicubic interpolation of a single member of the low resolution left and right views of the scene. These depth-map comparisons give two values with which to rate the accuracy of the 3D reconstruction, these are Average Error Per Depth (AEPD) and Number of Perfect Depths (NPD), these two values are the parallels to the AEPP and NPP values used

in [1]. Calculation of AEPD and NPD are shown below:

$$\frac{\sum_{j,k=1,1}^{MN}(|R_{jk}-T_{jk}|)}{M\,N} = \text{AEPD} \quad (6)$$

$$\frac{\sum_{j,k=1,1}^{MN}[(R_{jk}-T_{ijk})=0]}{M\,N} \times 100 = \text{NPD} \quad (7)$$

Where R is the reconstructed depth-map and T is the target depth-map, M and N are the width and height of the depth-map matrix. NPD is generated as a percentage of the number of pixels which are correct. The following tables show the results of applying an SDN to a number of stereo scenes and generating depth-map comparisons to the original views and the bicubic interpolation of a low resolution sample of each. In each case the most successful result is in bold.

TABLE I
COMPARATIVE SCORES FOR 3D RECONSTRUCTION

	Bicubic		SDN	
	NPD	**AEPD**	**NPD**	**AEPD**
A	58.37	2.99	**59.81**	**2.90**
B	55.13	3.20	**55.15**	**3.00**
C	57.11	3.06	**61.29**	**2.71**
D	**62.26**	**1.41**	53.28	1.69
E	57.31	3.09	**60.87**	**2.74**
F	58.36	2.94	**62.05**	**2.65**
G	**59.92**	**2.59**	58.83	2.82
H	**60.09**	**2.62**	58.92	2.80

TABLE II
COMPARATIVE SCORES FOR 3D RECONSTRUCTION WITH NOISE

	Bicubic		SDN	
	NPD	**AEPD**	**NPD**	**AEPD**
A	40.15	6.02	**48.43**	**4.37**
B	40.96	6.04	**47.35**	**4.42**
C	39.94	6.02	**48.94**	**4.32**
D	49.77	3.49	**53.69**	**2.10**
E	40.10	6.10	**48.75**	**4.37**
F	40.55	6.09	**49.39**	**4.27**
G	37.10	6.09	**46.19**	**4.40**
H	37.00	6.03	**45.96**	**4.39**

Table I shows that the SDN approach can produce a stereo pair which is at least as effective as bicubic interpolation when used in a disparity map generator. The same views subjected to noise are enhanced to a higher standard using the SDN due to its ability to suppress that noise, this can be seen in table II. In these tests drop-out noise with variance of 0.05 was added to all 6 frames before the application of any enhancement algorithm. There are two main reasons for the advantage the SDN approach has when compared with an interpolation approach. Firstly, the SDN is able to remove the additive noise which has been introduced into the low resolution frames. This allows many of the incorrect pixels to be removed before the disparity map is generated. If these are not removed they appear as differences and the disparity map generated contains errors, these can be seen in the bicubic result in Fig. 2(b) as vertical "spikes". Secondly, during enhancement the SDN attempts to retain as much original information as possible, whereas the bicubic interpolation modifies a pixel slightly to form a more visually appealing image. The downside of these visually appealing images is that they are generated slightly differently between the left and right view. This again causes the disparity map generator to detect these locations as differences and use them in the estimation of distances. Observing Fig. 2 it is clear that using an SDN to perform the enhancement of the left and right views was more beneficial to the 3D reconstruction algorithm than using bicubic interpolation. The levels in the reconstruction are more clearly defined within the SDN result and the number of "spikes" indicating points which have been incorrectly modeled are at a minimum.

V. CONCLUDING REMARKS

SDNs have been applied to the enhancement of 3D reconstructions of scenes. It has been shown that the use of SDNs to enhance the left and right views of a scene results in a 3D representation with fewer errors. The ability of the SDNs to remove noise was also demonstrated here when noise was independently added to the left and right streams before enhancement. This improvement in reconstruction quality is due to the different priorities of SDNs in comparison to bicubic interpolation. SDNs are trained to produce output pixels which are correct whereas bicubic output pixels aim to be "smooth" in comparison to their neighbours

Future Work

Pre-combination: In order to allow the use of existing trained SDNs stereo images can be combined to produce a single image to be processed by the network. The disadvantage of pre-combination is the loss of original information it causes.

b) Cyclopean Views: Each stereo pair will be merged in a pre-processing step to generate a cyclopean view. This is the image which corresponds to that which would have been captured by a camera placed between the two views. Algorithms exist which can generate the cyclopean view given the two stereo images. One of these is used and the image it produces is passed as the input to the enhancement systems.

c) Interlaced Views: Another method of pre-combination is based on interlacing the stereo pairs. A new image is generated using the even numbered rows from one image and the odd numbered lines from the other. This allows the complete image to be passed as one into the enhancement system. This provides the data from the two views but does reduce the resolution of the data in the dimension in which it is interlaced.

d) Depth Maps/3D Views: Depth maps from different times could be used to produce a depth map of higher resolution. It may be possible to increase the number of input and output neurons of the network to allow a 3D block of pixels to be passed in and out of the network. This should allow the same routing functionality as with the 2D inputs to

produce an output scene with more correct voxels (Volume Elements) than plain 3D interpolation.

REFERENCES

[1] L. E. Hibell, H. Liu, and D. J. Brown, "Combining multi-frame images for enhancement using self-delaying dynamic networks," in *IEEE World Congress on Computational Intelligence*, Canada, 2006.
[2] "Eye movements," 2004. [Online]. Available: www.cis.rit.edu/vpl/eye_movements.html
[3] J. Cooper, "All about strabismus," 2006. [Online]. Available: http://www.strabismus.org/ all_about_strabismus.html
[4] J. P. C. Southall, *Physiological Optics*. New York, NY: Dover, 1937.
[5] R. D. Henkel, "Fast stereovision by coherence detection," in *Computer Analysis of Images and Patterns*, 1997, pp. 297–304.
[6] ——, "A simple and fast neural network approach to stereo vision," in *NIPS'97 in Denver*. Cambridge: MIT Press, 1997, pp. 808–814.
[7] R. Tsai and T. Huang, *Multiframe Image Restoration and Registration*, R. Tsai and T. Huang, Eds. JAI Press, 1984, vol. 1.
[8] M. Irani and S. Peleg, "Super resolution from image sequences," *10th ICPR*, vol. 2, pp. 115–120, 1990.
[9] ——, "Improving resolution by image registration," *CVGIP: Graphical Models and Image Processing*, vol. 53, no. 3, pp. 231–239, May 1991.
[10] R. Schultz and R. Stevenson, "Extraction of high resolution frames from video sequences," *IEEE Transactions on Image Processing*, vol. 5, no. 6, pp. 996–1011, 1996.
[11] S. Kim, N. Bose, and H. Valenzuela, "Recursive reconstruction of high resolution image from noisy undersampled multiframes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 6, 1990.
[12] S. Borman and R. Stevenson, "Spatial resolution enhancement of low resolution image sequences. A comprehensive review with directions for future research," University of Notre Dame, Tech. Rep., 1998.
[13] E. Salari and S. Zhang, "Integrated recurrent neural network for image resolution enhancement from multiple image frames," in *IEE Proceedings on Vision, Image and Signal Processing*, vol. 150, no. 5, 2003.
[14] C. Miravet and F. B. Rodríguez, "A hybrid MLP-PNN architecture for fast image superresolution." in *ICANN*, 2003, pp. 417–424.
[15] ——, "Accurate and robust image superresolution by neural processing of local image representations." in *ICANN (1)*, 2005, pp. 499–505.
[16] M. Negnevitsky, *Artificial Intelligence*, 2nd ed. Harlow, UK: Pearson, 2005.
[17] A. S. Ogale and Y. Aloimonos, "A roadmap to the integration of early visual modules," *International Journal of Computer Vision: Special Issue Of Early Cognitive Vision*, In Press.
[18] "Microsoft Research, Cambridge. i2i: 3D visual communication," March 2006. [Online]. Available: http://research.microsoft.com/vision/cambridge/i2i/default.htm