

NEURAL NETWORK APPROACH TO BAYESIAN BACKGROUND MODELING FOR VIDEO OBJECT SEGMENTATION

Dubravko Čulibrk, Oge Marques, Daniel Socek, Hari Kalva and Borko Furht
Department of Computer Science and Engineering
Florida Atlantic University, 777 Glades Rd., Boca Raton, FL 33431, USA
{dculibrk,omarques,dsocek}@fau.edu, {hari,borko}@cse.fau.edu

Keywords: Video, Object segmentation, Background modelling, Bayesian modelling, Neural Networks.

Abstract: Object segmentation from a video stream is an essential task in video processing and forms the foundation of scene understanding, object-based video encoding (e.g. MPEG4), various surveillance and 2D to pseudo 3D conversion applications. The task is difficult and exacerbated by the advances in video capture and storage (e.g. HDTV, QuadHDTV). Increased resolution of the sequences requires development of new, more efficient algorithms for object detection and segmentation. The paper presents a novel neural network based approach to background modelling for motion based object segmentation in video sequences. The proposed approach is designed to enable efficient, highly-parallelized hardware implementation. Such a system would be able to achieve real time segmentation of high-resolution sequences.

1 INTRODUCTION

Object detection and segmentation from a video stream are essential tasks in video processing and form the foundation of scene understanding, object-based video encoding (e.g. MPEG4), various surveillance applications, as well as the emerging research into 2D to pseudo 3D video conversion. The task is difficult and exacerbated by the advances in video capture and storage (e.g. HDTV, QuadHDTV). Increased complexity of the sequences requires development of new, more efficient algorithms for object detection and segmentation.

Commonly used approach to extract foreground objects from the image sequence is through background suppression [39-41], when the video is grabbed from a stationary camera. However, the task becomes difficult when the background contains shadows and moving objects, and undergoes illumination changes. Significant scientific effort has been spent on the development of adaptive models of background and segmentation techniques. A number of proposed techniques are able to achieve real-time processing of comparatively small video formats (e.g. 120x160 pixels, CIF resolution) and, usually, at somewhat reduced frame rates. It is unlikely, however, that the existent object detection

approaches will be able to efficiently cope with the increase in the resolution of video sequences. The development of an object detection approach, which would allow for efficient hardware implementation and object detection in real-time for high-complexity video sequences (in terms of the frame size as well as background changes), is the focus of this paper. A new neural network structure is proposed, to serve both as an adaptive Bayesian model of the background in a video sequence and an algorithm for background subtraction and foreground object detection and segmentation.

The rest of the paper is organized as follows: Next section provides a survey of related published work. The third section describes the main aspects of the proposed approach. The fourth is dedicated to the presentation of simulation results. The last section holds the conclusions and some directions for future work.

2 RELATED WORK

Some of the early object segmentation methods dealing with the instances of non-stationary background were based on smoothing the colour of a background pixel over time using different filtering

techniques such as Kalman filters[21][22], or Gabor filters [20] to create a reference background frame. The reference frame is a model of background, which is constantly updated and used to segment the foreground objects by subtracting it from the current frame of the input sequence. However, since these methods are based on the most restrictive assumption that movements of the background are much slower than those of the objects to be segmented, they are not particularly effective for sequences with high-frequency background changes.

Slightly better results were reported for techniques that rely on a Gaussian -based statistical model whose parameters are recursively updated in order to follow gradual background changes within the video sequence[23]. More recently, this model was significantly improved by employing a Mixture of Gaussians (MoG), where the values of the pixels from background objects are described by multiple Gaussian distributions[24][26][27]. This model was considered promising since it showed good foreground object segmentation results for many outdoor sequences. However, weaker results were reported [28] for video sequences containing non-periodical background changes (e.g. due to waves and water surface illumination, cloud shadows, and similar phenomena). These models are parametric in the sense that they incorporate underlying assumptions about the PDFs they are trying to estimate.

In 2003, Li et al. proposed a method for foreground object detection employing a Bayes decision framework [28][29]. The method has shown promising experimental object segmentation results even for the sequences containing complex variations and non-periodical movements in the background. In addition to the generic nature of the algorithm where no a priori assumptions about the scene are necessary, the authors claim that their algorithm can handle a throughput of about 15 fps for CIF video resolution. The approach is specific in the fact that it uses a statistical model of for the changes between the current frame and the reference background image maintained by applying an Infinite Impulse Response (IIR) filter to the sequence. A Bayesian classifier is then used to classify the changes, detected through frame differencing between the current frame and the reference frame, as pertinent to background objects or foreground objects. The statistical model is non-parametric since it does not impose any specific shape to the PDFs learned. However, for reasons of efficiency and improving results the authors applied binning of the features and assigned single probability to each bin, leading to a discrete representation of PDFs. The model is general in terms of features extracted from the sequence and

they experimented with the use of different features. The results of these experiments are reported in [29].

Recently the approach of Li et al. has been adopted and extended to create a part of a surveillance system intended for maritime environments [30]. The results in this domain have been improved by altering the frame differencing step of the algorithm as well as using a color-based still image segmentation instead of the morphological operations in the post-processing of the background-subtraction results.

While the use of Bayesian models as basis for background subtraction is not new, it has been limited by the fact that they are general in the sense that they impose no constraints on the shape of the estimated probability density function. This typically makes them more computationally expensive than most of their more restrictive counterparts (e.g.[23][24][26][27]). However, moving away from the particle estimator systems used typically to estimate probability density functions in the Bayesian models [29-30] to neural networks, it is possible to make them suitable for parallel execution and increase their effectiveness.

Classical Probabilistic Neural Network (PNN) architecture can be used to create an efficient[1][4][8], but this extant solution is a supervised learning classifier and as such unable to cope with the task of background subtraction without a supervisor classifier.

In [1] authors present an unsupervised video object (VO) segmentation and tracking algorithm based on an adaptable neural-network architecture. The proposed scheme comprises a VO tracking module and an initial VO estimation module. Object tracking is handled as a classification problem and implemented through an adaptive network classifier, which, however, relies on the results of the initial video object segmentation module to adjust itself to the variations of the sequence. Based on the video object segmentation results a set is constructed, which is used to retrain the network, as proposed by the same authors in [3]. To determine when the network should be retrained a decision mechanism is used. It consists of a shot cut detection module and an operational environment change module. The first is based on the principle that all different poses that a VO takes within a shot are usually strongly correlated to each other, while the second is incorporated as a safety valve to confront gradual but significant content changes within a shot. To detect shot transitions an approach proposed in [2] is used while the gradual changes in the environment are estimated based on the error of the neural network with respect to the results achieved by the initial object segmentation algorithm. To this end the authors extract features of the objects segmented by

the neural network and compare them to the features of the initially segmented objects. The accuracy of the decision module is crucial for the performance of the system as a whole, since the retraining of the network is computationally expensive and frequent retraining ruins the computational efficiency of the algorithm, while not retraining when needed leads to poor classification results. The authors claim improved performance of their approach over the conventional motion-based tracking algorithms. Although the whole segmentation algorithm is an unsupervised learner, clearly the retrainable neural network is itself a supervised learner, differing from the approach proposed here.

An approach employing a Probabilistic Neural Network (PNN) classifier in a time varying environment is proposed in [4] [5]. A PNN was used to classify clouds based on their spectral and temperature features in the visible and infrared GOES 8 (Geostationary Operational Environmental Satellite) imagery data. A temporal updating approach for the PNN was developed to increase the classification accuracy by accounting for the temporal changes in the data. The adaptation of the PNN is supervised by Markov chain models of the temporal contextual information combined with the mixture of Gaussians (MoG) maximum likelihood estimation. The network itself is a supervised learner and is updated every time a new frame is processed.

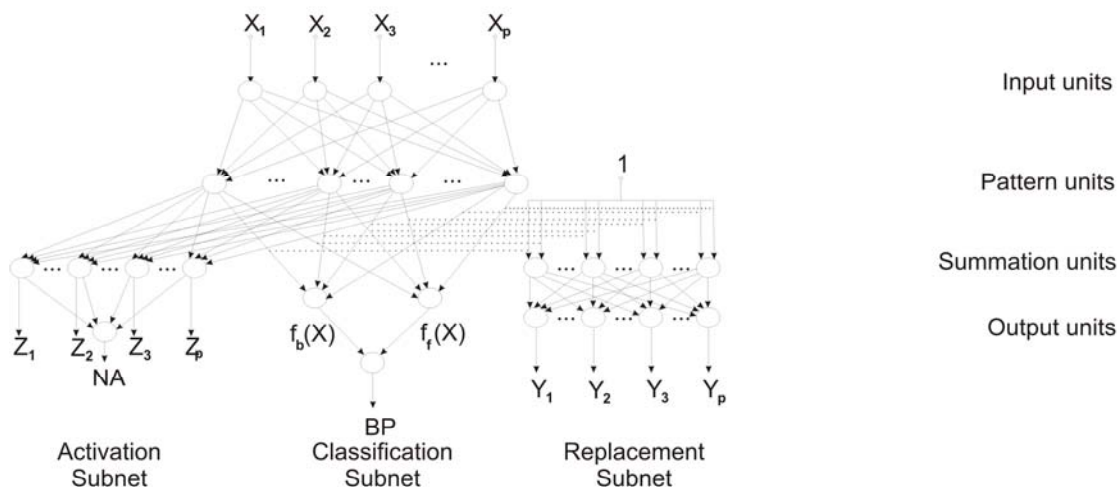
BACKGROUND MODELLING NEURAL NETWORK (BNN)

The proposed background modelling and subtraction approach relies on a novel adaptive neural network. The proposed architecture employs an adapted General Regression Neural Network (GRNN) [8][9] component, to serve as an estimator of the density function of probability of certain features belonging to background. GRNNs, typically used as Bayesian classifiers, are supervised classifiers, requiring a training set. However, in the domain of background modelling it was possible to extend them to form new neural network architecture which is an unsupervised learner. This Background Modelling Neural Network (BNN) is suitable to serve both as a statistical model of the background at each pixel in the video sequences and highly parallelized background subtraction algorithm.

The design of BNN relies on a basic background modelling idea: *feature values corresponding to background object will occur most of the time, i.e. more often than those pertinent to the foreground.*

Three tasks, typical for probabilistic background modelling [MoG][Li], which BNN should perform have been identified:

1. Storing the values of the features and learning the probability with which each value corresponds to background / foreground,



2. Determining the state in which new feature values should be introduced in to the model (i.e. when the statistics already learned are insufficient to make a decision),

3. Determining which stored feature value should be replaced with the new values.

The two latter requirements are consequences of the fact that real systems are limited in terms of the number of feature values that can be stored to achieve efficient performance.

The structure of BNN, shown in Figure 1, has three distinct subnets. The classification subnet is a GRNN [9]. It is a central part of BNN concerned with approximating the Probability Density Function (PDF) of pixel feature values belonging to background/foreground. The GRNN is a neural network implementation of a Parzen estimator [11]. This class of PDF estimators asymptotically approaches the underlying parent density, provided that it is smooth and continuous.

The classification subnet contains 3 layers of neurons. Input neurons of this net simply map the inputs of the network, which are the values of the features for a specific pixel. The output of the pattern neurons is a nonlinear function of Euclidean

distance between the input of the network and the stored pattern for that specific neuron. The nonlinear function used is as proposed by Parzen. The only parameter of this subnet is a so called smoothing parameter used to determine the shape of the nonlinear function. The structure of a pattern

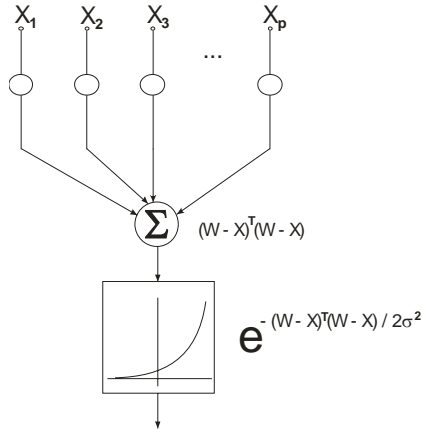


Figure 2 Pattern neuron of GRNN. neuron is shown in Figure 2.

The output of the summation units of the classification subnet is the sum of their inputs. The subnet has two summation neurons: one to calculate the probability of pixel values belonging to background and the other for calculating the probability of belonging to foreground.

The classification subnet requires no training to store the patterns (feature values) representative of background. This is accomplished simply by setting the weights of the connections between the input and pattern neurons to the value of the features of the pattern to be stored. The classification subnet diverges from GRNN in the way the weights between the pattern and summation neurons are determined. These values are used to store the confidence with which a pattern belongs to the background/foreground. The weights of these connections are updated with each new value of a pixel at a certain position received (i.e. with each frame), according to the following recursive equations:

$$W_{if}^{t+1} = (1 - \beta)W_{if}^t + \beta \quad (1)$$

$$(2)$$

$$W_{ib}^{t+1} = (1 - \beta)W_{ib}^t + \beta \quad W_{if}^{t+1} = (1 - \beta)W_{if}^t$$

when the maximum response is that of the i -th neuron, and

$$(3)$$

$$(4)$$

if the maximum response is not that of the j -th neuron, where:

W_{ib}^t - value of the weight between the i -th pattern neuron and the background summation neuron,

W_{if}^t - value of the weight between the i -th pattern neuron and the foreground summation neuron,

β - learning rate.

Equations 1-4, express the notion that whenever an instance pertinent to a pattern neuron is encountered, the probability that that pattern neuron is activated by a feature value vector belonging to the background, is increased. Naturally, if that is the case the, probability that the pattern neuron is excited by a pattern belonging to foreground is decreased. Vice versa, the more seldom a feature vector value corresponding to a pattern neuron is encountered the more likely it is that the patterns represented by it belong to foreground objects. By adjusting the learning rates, it is possible to control the speed of the learning process.

The output of the classification subnet indicates whether the output of the background summation neuron is higher than that of the foreground summation neuron, i.e. that it is more probable that

the input feature value is due to a background object rather than a foreground object.

The activation and replacement subnets are Winner-Take-All (WTA) neural networks. A WTA network is a parallel and fast way to determine minimum or the maximum of a set of values, consistent with the task of doing so within a neural-network based solution. In particular, these subnets are extensions of one-layer feedforward MAXNET (1LF-MAXNET) proposed in [37].

The activation subnet performs a dual function: it determines which of the neurons of the network has maximum activation (output) and whether that value exceeds a threshold $replace_crit = W_{ib}^t + |W_{ib}^t - W_{if}^t|$ provided as a parameter to the algorithm. If it does not, the BNN is considered inactive and replacement of a pattern neuron's weights with the values of the current input vector is required. If this is the case, the feature is considered to belong to a foreground object.

The first layer of this network has the structure of a 1LF-MAXNET network and a single neuron is used to indicate whether the network is active. The output of the neurons of the first layer of the network can be expressed in the form of the following equation:

$$Y_j = X_j \cdot \prod_{i=1}^p \{F(X_j - X_i) \mid i \neq j\} \quad (5)$$

where:

The output of the first layer of the activation subnet will differ from 0 only for the neurons with maximum activation and will be equal to the maximum activation. In Figure 1 these outputs are indicated with $Z_1 \dots Z_p$. Figure 3 shows the inner structure of a neuron in the first layer of the subnet.

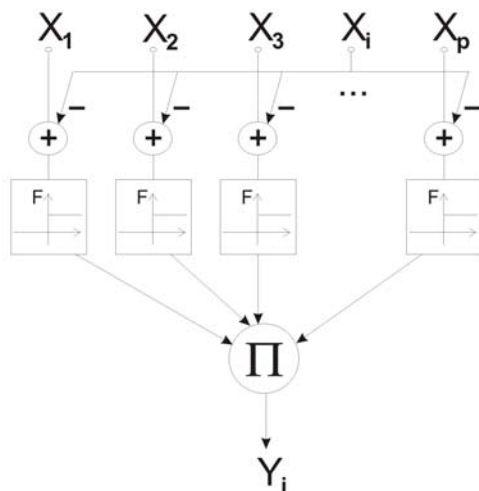


Figure 3 Structure of processing neurons of the activation subnet.

A single neuron in the second layer of the activation subnet is concerned with detecting whether the BNN is active or not and its function can be expressed in the form of the following equations:

$$(6)$$

where:

and θ is the activation threshold, which is provided to the network as a parameter.

Finally, the replacement subnet in figure 1 can be viewed as a separate neural net with the unit input. However, it is inextricably related to the classification subnet since each of the replacement subnet first-layer neurons is connected with the input via synapses that have the same weight as the two output synapses between the pattern and summation neurons of the classification subnet. Each pattern neuron has a corresponding neuron in the replacement net. The function of the replacement net is to determine the pattern neuron that minimizes the criterion for replacement, expressed by the following equation:

$$(7)$$

The criterion is a mathematical expression of the idea that those patterns that are least likely to belong to the background and those that provide least confidence to make the decision should be eliminated from the model.

The neurons of the first layer calculate the negated value of the replacement criterion for the pattern neuron they correspond to. The second layer is a 1LF-MAXNET that yields non-zero output corresponding to the pattern neuron to be replaced.

To form a complete background-subtraction solution a single instance of a BNN is used to model the features at each pixel of the image.

EXPERIMENTS AND RESULTS

The approach is intended to serve as basis for the design of a hardware component, which would be able to exploit its highly parallel nature. However, as proof of concept, a simulation application, which can be run on a typical PC, has been developed. While this simulation is sequential in its execution and cannot provide a valid estimate of the speed of the target hardware system, it can demonstrate the segmentation ability of the system.

In a hardware implementation the delay of the network corresponds to the time needed by the signal to propagate through the network and time

required to update it. In a typical FPGA implementation this can be done in less than 20 clock cycles, which corresponds to a 2ms delay through the network, for a FPGA core running at 100ns clock rate.

The simulation application implements BNNs containing 20 processing, two summation and one output neuron per pixel in the classification subnet. The activation and replacement subnet attribute for additional 20, i.e. 41 processing units respectively, bringing up the total of neurons used per pixel to 84. The input neurons of the classification shown in Figure 1 just map the input to the output and need not be implemented as such.

The simulation is capable of processing a single frame of size 720x480 in 2.25 seconds on average, which translates to 8 frames of 160x120 pixels per second or 2.2 frames per second(fps) for images sized 320x240 pixels.

The primary sequence used for testing is a maritime environment sequence containing 18230 frames of 720x480 pixels, corresponding to a bit more than 10 minutes of recording at 30 frames per

second. It contains a large number of diverse vessels that the algorithm tries to segment and is complex in terms of background changes related to the water-surface. Two consecutive frames from the sequence as well as the results of segmentation are given in Figure 3. Coloured pixels correspond to the foreground. Green are classified as foreground due to the BNNs recognizing that these are new values not yet stored, while the red ones are stored but classified as foreground based on the learned PDFs. No morphological operations, typically used to remove spurious one pixel effects and make the object solid, have been performed on the segmentation images shown in Figure 3. These are currently performed as a post-processing step, but will ultimately be implemented as a neural network.

The learning rate of the networks was set to 0.005 and the smoothing parameter for the classification subnet used was set to 10. The activation threshold of the activation subnet was set to 0.95.



Figure 4 Two consecutive frames from a test sequence (top) and the segmentation result (bottom).

CONCLUSION AND FURTHER RESEARCH

A new motion based object segmentation and background modelling approach has been proposed. The basis of the approach is employment of a novel neural network architecture designed specifically to serve as a model of background in video sequences and a Bayesian classifier to be used for object segmentation. The new Background Modelling Neural Network is an unsupervised classifier. The proposed model is independent of the features used.

The design is intended to be implemented in hardware, allowing for highly-parallelized execution.

We present results of the simulation of the system on a PC, using a complex maritime sequence. The simulation itself allows for a fairly fast segmentation of objects.

Future work will focus on the use of features different than RGB values, development of a FPGA based system and enhancing the segmentation results through the use of cues not related to motion.

REFERENCES

- Haritaoglu, I., Harwood, D., and Davis, L., 2000. W⁴ Real-time surveillance of people and their activities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 809–830.
- Stauffer, C. and Grimson, W. 2000. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 747–757.
- K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Proceedings of IEEE Int'l Conf. On Computer Vision*, pp. 255–261. 1999.
- K. P. Karmann, A. von Brandt, “A.Moving object recognition using an adaptive background memory”, *Time-varying Image Processing and Moving Object Recognition*, 2, pp. 297-307, Elsevier Publishers B.V., Amsterdam, 1990.
- C. Ridder, O. Munkelt, H. Kirchner, “Adaptive background estimation and foreground detection using Kalman-filtering”, in *Proc. of International Conference on Recent Advances in Mechatronics (ICRAM'95)*, pp. 193-199, 1995.
- Jain, A.K., Ratha, N.K., Lakshmanan, S., “Object detection using Gabor filters”, *Journal of Pattern Recognition*, vol. 30, pp. 295-309, 1997.
- T.E. Boulton, R. Micheals, X.Gao, P. Lewis, C. Power, W. Yin, A. Erkan: “Frame-rate omnidirectional surveillance and tracking of camouflaged and occluded targets”, in *Proc. of IEEE Workshop on Visual Surveillance*, pp. 48–55, 1999.
- T.J. Ellis, M. Xu, “Object detection and tracking in an open and dynamic world”, in *Proc. of the Second IEEE International Workshop on Performance Evaluation on Tracking and Surveillance (PETS'01)*, 2001.
- C. Stauffer, W.E.L. Grimson, “Learning patterns of activity using real-time tracking”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 747-757, 2000.
- L. Ya, A. Haizhou, X. Guangyou, “Moving object detection and tracking based on background subtraction”, in *Proc. of SPIE Object Detection, Classification, and Tracking Technologies*, pp. 62-66, 2001.
- L. Li, W. Huang, I.Y.H. Gu, Q. Tian, “Foreground object detection from videos containing complex background”, in *Proc. of the Eleventh ACM International Conference on Multimedia (MULTIMEDIA'03)*, pp. 2-10, 2003.
- L. Li, W. Huang, I.Y.H. Gu, Q. Tian, “Statistical Modeling of Complex Backgrounds for Foreground Object Detection”, *IEEE Trans. Image Processing*, vol. 13, pp. 1459-1472, Nov. 2004.
- D. Socek, D. Culibrk, O. Marques, H. Kalva, B. Furht, “A Hybrid Color-Based Foreground Object Detection Method for Automated Marine Surveillance”, in *Proc. of the Advanced Concepts for Intelligent Vision Systems Conference (ACIVS 2005)*, 2005. (in print)
- T.E. Boulton, R. Micheals, X.Gao, P. Lewis, C. Power, W. Yin, A. Erkan: “Frame-rate omnidirectional surveillance and tracking of camouflaged and occluded targets”, in *Proc. of IEEE Workshop on Visual Surveillance*, pp. 48–55, 1999.
- A. Doulamis, N. Doulamis, K. Ntalianis, and S. Kollias, “An Efficient Fully Unsupervised Video Object Segmentation Scheme Using an Adaptive Neural-Network Classifier Architecture”, *IEEE Trans. On Neural Networks*, vol. 14, pp. 616-630, May 2003.
- B. Tian, M. R. Azimi-Sadjadi, T. H. Vonder Haar, and D. Reinke, “Temporal Updating Scheme for Probabilistic Neural Network with Application to Satellite Cloud Classification,” *IEEE Trans. Neural Networks*, vol. 11, pp. 903–920, July 2000.
- D. F. Specht, “Probabilistic neural networks,” *Neural Netw.*, vol. 3, pp. 109–118, 1990.
- A. Doulamis, N. Doulamis, and S. Kollias, “On line retrainable neural networks: Improving the performance of neural networks in image analysis problems,” *IEEE Trans. Neural Networks*, vol. 11, Jan. 2000.
- M. R. Azimi-Sadjadi, W. Gao, T. H. Vonder Haar, and D. Reinke, “Temporal Updating Scheme for Probabilistic Neural Network With Application to Satellite Cloud Classification—Further Results,” *IEEE Trans. Neural Networks*, vol. 12, pp. 1196–1203, September 2001.

- D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Networks*, pp. 568–576, Mar. 1991.
- E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, Vol. 33, pp. 1065-1076, Sept. 1962.
- H. K. Kwan, "One-layer feedforward neural network fast maximum/minimum determination," *Electronics Letters*, pp. 1583-1585, 1992.