A NEURAL NETWORK APPROACH TO BAYESIAN BACKGROUND MODELING FOR VIDEO OBJECT SEGMENTATION

Dubravko Culibrk*, Oge Marques+, Daniel Socek†
Department of Computer Science and Engineering, Florida Atlantic University
Boca Raton FL 33431, USA
dculibrk@fau.edu, *omarques@fau.edu*+, *dsocek@fau.edu*†

Hari Kalva×, Borko Furht‡
Department of Computer Science and Engineering, Florida Atlantic University
Boca Raton FL 33431, USA
hari@cse.fau.edu×, *borko@cse.fau.edu*‡

Keywords: Video processing, Object segmentation, Background modeling, Bayesian modeling, 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), and various surveillance and 2D-to-pseudo-3D conversion applications. The task is difficult and exacerbated by the advances in video capture and storage. 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 modeling 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. 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 (Haritaoglu et al., 2000)(Stauffer and Grimson, 2000), 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. 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 a parallelized 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.

The proposed solution employs a feed-forward neural network to achieve background subtraction. To this end, a new neural network structure is designed, serving 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. Neural networks posses intrinsic parallelism which can be exploited in a suitable hardware implementation to achieve fast segmentation of foreground objects.

The rest of the paper is organized as follows: Section 2 provides a survey of related published work. Section 3 describes the main aspects of the proposed approach. Section 4 is dedicated to the presentation of simulation results. Section 5 contains 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 color of a background pixel over time using different filtering techniques such as Kalman filters (Karmann and von Brandt,

1990), or Gabor filters (Jain et al., 1997) to create a reference background frame. It is used to segment the foreground objects by subtracting the reference frame from the current frame of the input sequence. However, these methods 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(Boult et al., 1999). 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(Stauffer and Grimson, 2000). This model was considered promising since it showed good foreground object segmentation results for many outdoor sequences. However, weaker results were reported (Li et al., 2004) 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 probability density functions (PDFs) they are trying to estimate.

In 2003, Li et al. proposed a method for foreground object detection employing a Bayes decision framework (Li et al., 2004). The method has shown promising experimental object segmentation results for 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. 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 (Li et al., 2004).

Recently the approach of Li et al. has been adopted and extended to create a part of a surveillance system intended for maritime environments (Socek et al., 2005).

While the use of Bayesian models as bases 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.(Boult et al., 1999)(Stauffer and Grimson, 2000)). However, moving away from the particle estimator systems used typically to estimate probability density functions in the Bayesian models (Li et al., 2004) to neural networks, it is possible to make them suitable for parallel execution and increase their effectiveness.

Classical Probabilistic Neural Network (PNN)(Specht, 1990) architecture has been used by researchers to improve the object segmentation(Doulamis et al., 2003) and perform the classification of segmented objects (Azimi-Sadjadi et al., 2001). In both solutions the neural network is a supervised learning classifier guided by a a different supervisor classifier algorithm.

In (Doulamis et al., 2003) authors present an unsupervised video object (VO) segmentation and tracking algorithm based on an adaptive neural-network architecture. 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. The neural network is but a part of the background segmentation algorithm. Hence, the whole system does not posses inherent parallelism of the PNN. As such, the system is not suitable to serve as basis of an efficient hardware implementation.

An approach employing a PNN classifier in a time varying environment is proposed in (Azimi-Sadjadi et al., 2001). 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 network itself is a supervised leaner and is updated every time a new frame is processed. As in the approach of (Doulamis et al., 2003), the PNN is a submodule of the system, and its parallelism can only partially be exploited in a hardware implementation.

3 BACKGROUND MODELING NEURAL NETWORK (BNN)

The proposed background modeling and subtraction approach relies on a novel adaptive neural network. The architecture employs an adapted General Regression Neural Network (GRNN) (Specht, 1991) component, to serve as an estimator of the PDF of certain features belonging to background. GRNNs, typically used as Bayesian classifiers, are supervised clas-

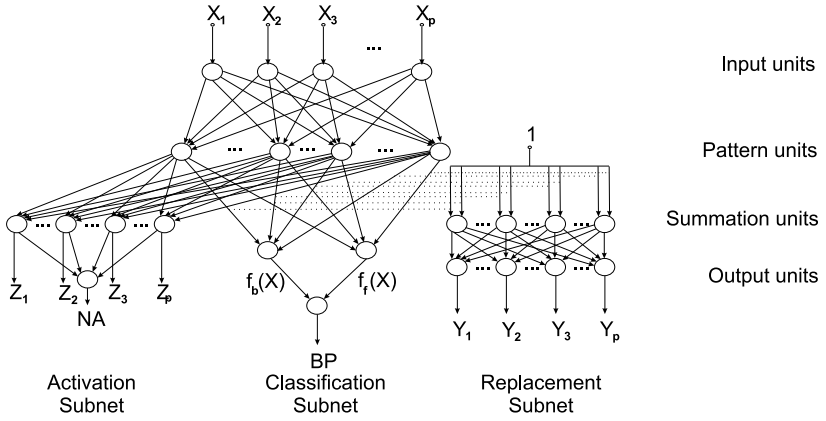


Figure 1: Structure of Background Modeling Neural Network.

sifiers, requiring a training set. However, in the domain of background modeling it was possible to extend them to form new neural network architecture which is an unsupervised learner. This Background Modeling Neural Network (BNN) is suitable to serve both as a statistical model of the background at each pixel position in the video sequences and highly parallelized background subtraction algorithm. The design of BNN relies on a basic background modeling 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 modeling (Stauffer and Grimson, 2000)(Li et al., 2004), 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 into 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 (Specht, 1991). It is a central part of BNN concerned with approximating the PDF of pixel feature values belonging to background/foreground. The GRNN is a neural network implementation of a Parzen estimator (Parzen, 1962). This class of PDF estimators asymptotically approaches the underlying parent density, provided that it is smooth and continuous.

The classification subnet contains three layers of neurons. Input neurons of this network 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 (σ) used to determine the shape of the nonlinear function. The structure of a pattern 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.

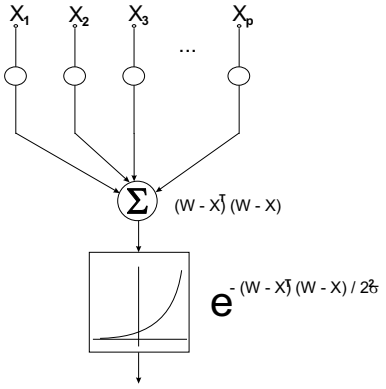


Figure 2: Pattern neuron of GRNN.

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:

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

$$W_{ib}^{t+1} = (1 - \beta) * W_{ib}^{t} + \beta \quad (2)$$

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

$$W_{if}^{t+1} = (1 - \beta) * W_{if}^{t} + \beta \quad (3)$$

$$W_{ib}^{t+1} = (1 - \beta) * W_{ib}^{t} \quad (4)$$

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 at time t,
- W_{if}^{t} - value of the weight between the i-th pattern neuron and the foreground summation neuron at time t,
- 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 (Kwan, 1992).

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 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:

$$Y_j = X_j \times \prod_{i=1}^{P} \{F(X_j - X_i | i \neq j)\} \quad (5)$$

where:

$$F(z) = \begin{cases} 1, & \text{if } z \geq 0; \\ 0, & \text{if } z < 0; \end{cases} \quad (6)$$

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 $Z_1, , Z_P$. 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:

$$NA = F(\sum_{i=1}^{P} Z_i - \theta) \quad (7)$$

where F is given by Equation 6 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:

$$replacement\_criterion = W_{if}^{t} + |W_{ib}^{t} - W_{if}^{T}| \quad (8)$$

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.

(a) (b) (c)

Figure 3: Results obtained for representative frames: (a) frame from the original sequence, (b) segmentation results obtained for the frame shown, (c) segmentation result with morphological enhancement.

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.

4 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, to evaluate the approach, 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.

The approach has been tested on a number of maritime environment sequences. All of the sequences used in the experiments are currently available at the following link: *http://sasi.cse.fau.edu/ccst/sequences*. Primary sequence used for testing is that of a port inlet, taken by a static camera. The sequence consists of 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, in terms both of color and size, moving at different rates, in different directions and at different distance from the camera. The sequence is also complex in terms of background changes related to the water-surface.

A representative frame from the sequence as well as the result of segmentation are given in Figure 3. Grey pixels correspond to the foreground. Figure 3(c) shows the same segmentation result when a morphological open and then morphological close operation are applied. The support region for the operations is a two pixel wide square.

A detail of the first frame of the original sequence (shown in figure 3(a)) containing several small objects as well as the segmentation result for that part of the frame with and without morphological operations applied is shown enlarged in Figure 4. Light grey pixels are classified as foreground due to the BNNs recognizing that these are new values not yet stored, while the dark grey ones are stored but classified as foreground based on the learned PDFs.

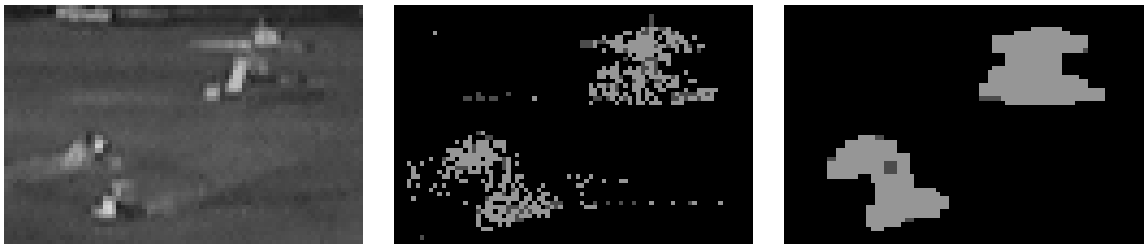
The neural networks used in the experiments are fairly simple. The simulation application implements BNNs containing 20 processing, two summation and one output neuron per pixel in the classification subnet. The activation and replacement subnet attribute for additional 20, i.e. 41 processing units respectively, 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 learning rate (β) of the networks was set to 0.005 and the smoothing parameter (σ) for the classification subnet used was set to 10. The activation threshold (θ) of the activation subnet was set to 0.95.

The performance of the simulation application allows for efficient experimenting. It is capable of processing a single frame of size 720×480 in 2.25 seconds on average, which translates to 8 frames of 160×120 pixels per second or 2.2 frames per second (fps) for images sized 320×240 pixels, on a 3.0 GHz Pentium IV based system.

5 CONCLUSION AND FURTHER RESEARCH

Object segmentation is a fundamental task in several important domains of video processing. The complexity of captured and stored video material is on the rise and current motion based segmentation algorithms are not capable of handling high-resolution sequences in real time. The possibility of resolving this problem through a highly parallelized approach is the



(a) (b) (c)

Figure 4: Details of: (a) a frame of the original sequence, (b) segmentation result and (c) segmentation result enhanced using morphological operations.

focus of the research presented.

The research resulted in a new motion based object segmentation and background modeling algorithm, proposed here. It is parallelized at sub-pixel level. 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 Modeling Neural Network is an unsupervised classifier, differing from the approaches published before. The proposed model is independent of the features used and general since it does not impose restrictions in terms of the probability density functions estimated.

A PC based system has been developed to evaluate the algorithm using a complex maritime sequence. The results obtained through these experiments are illustrated in the paper via representative frames.

Full exploitation of the algorithm's parallelism can be achieved only if the system is implemented in hardware, allowing for highly-parallelized execution.

Future work will proceed in several directions: Use of features different than RGB values will be explored to evaluate the impact of the choice of features on the performance of the system. Methods to enhance the segmentation, other than morphological transformations, will be explored (e.g. single frame color-based segmentation, depth cues from stereo sequences). Finally, development of a FPGA based system which would achieve real time segmentation of HDTV and QuadHDTV sequences will be explored.

REFERENCES

Azimi-Sadjadi, M. R., Gao, W., Haar, T. H. V., and Reinke, D. (2001). Temporal updating scheme for probabilistic neural network with application to satellite cloud classification-further results. In *IEEE Trans. Neural Networks, vol. 12, pp. 1196-1203*.

Boult, T., Micheals, R., X.Gao, Lewis, P., Power, C., Yin, W., and Erkan, A. (1999). Frame-rate omnidirectional surveillance and tracking of camouflaged and occluded targets. In *Proc. of IEEE Workshop on Visual Surveillance, pp. 48-55*.

Doulamis, A., Doulamis, N., Ntalianis, K., and Kollias, S. (2003). An efficient fully unsupervised video object segmentation scheme using an adaptive neural-network classifier architecture. In *IEEE Trans. On Neural Networks, vol. 14, pp. 616-630*.

Haritaoglu, I., Harwood, D., and Davis, L. (2000). W4 real-time surveillance of people and their activities. In *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, pp. 809-830*.

Jain, A., Ratha, N., and Lakshmanan, S. (1997). Object detection using gabor filters. In *Journal of Pattern Recognition, vol. 30, pp. 295-309*.

Karmann, K. P. and von Brandt, A. (1990). Moving object recognition using an adaptive background memory. In *Timevarying Image Processing and Moving Object Recognition, 2, pp. 297-307*. Elsevier Publishers B.V.

Kwan, H. K. (1992). One-layer feedforward neural network fast maximum/minimum determination. In *Electronics Letters, pp. 1583-1585*.

Li, L., Huang, W., Gu, I., and Tian, Q. (2004). Statistical modeling of complex backgrounds for foreground object detection. In *IEEE Trans. Image Processing, vol. 13, pp. 1459-1472*.

Parzen, E. (1962). On estimation of a probability density function and mode. In *Ann. Math. Stat., Vol. 33, pp. 1065-1076*.

Socek, D., Culibrk, D., Marques, O., Kalva, H., and Furht, B. (2005). 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)*.

Specht, D. F. (1990). Probabilistic neural networks. In *Neural Networks, vol. 3, pp. 109-118*.

Specht, D. F. (1991). A general regression neural network. In *IEEE Trans. Neural Networks, pp. 568-576*.

Stauffer, C. and Grimson, W. (2000). Learning patterns of activity using real-time tracking. In *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, pp. 747-757*.