ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

An ART-type network approach for video object detection

R.M. Luque, E. Domínguez, E.J. Palomo, J. Muñoz

Universidad de Malaga - Dept of Computer Science
Campus Teatinos s/n, 29071 - Malaga, Spain
{rmluque,enriqued,ejpalomo,munozp}@lcc.uma.es

Abstract.

This paper presents an ART-type network (adaptive resonant theory) to detect objects in a video sequence classifying the pixels as foreground or background. The proposed ART network (ART+) not only possesses the structure and learning ability of an ART-based network, but also uses a neural merging process to adapt the variability of the input data (pixels) in the scene along the time. Experimental results demonstrate the effectiveness and feasibility of the proposed ART+ approach for object detection. Standard datasets are used to compare the efficiency of the proposed approach against other traditional methods based on gaussian models.

1 Introduction

A visual surveillance system starts with motion detection, which aims at segmenting regions corresponding to moving objects from the rest of an image. Unfavourable factors, such as illumination variance, shadows and shaking branches, bring many difficulties to the acquisition and updating of background images.

There are many algorithms for resolving these problems including temporal average of an image sequence [1], adaptive Gaussian estimation [2],parameter estimation based on pixel processes [3], etc. Haritaoglu et al.[4] build a statistical model by representing each pixel with three values: the minimum and maximum intensity values, and the maximum intensity difference between consecutive frames observed during the training period. These three values are updated periodically.

Adaptive resonant theory (ART) was introduced by Grossberg [5] and thereafter different ART-type networks were subsequently developed by Carpenter et al. [6]. ART architectures are neural networks that carry out stable self-organisation for arbitrary sequences of input patterns. Without preliminary training, ART networks not only allow cluster centres to adapt to current conditions, but also allow on-line creation of new categories during classification phases. These characteristics are very useful for detecting motion in video sequences. The similarity measurement is further examined by the vigilance test to determine whether the input data belongs to the existing clusters or becomes the centre of a new cluster.

The background of a scene is characterised by the most activated clusters, since the background pixels are the most frequent in a scene. Clusters are labelled as background or foreground depending on their activity. In this sense, the proposed ART network (ART+) is capable of classifying the pixels of a scene according to this criterion, in which all pixels belong to the background clusters are classified as background.

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

The problem of determining optimal vigilance threshold is also introduced and is a novelty of our approach. This value determines the number of clusters and is updated along the time due to illumination changes and shadows in the video sequence. Moreover, a merging process is proposed to determine the optimal number of clusters and the optimal location of the cluster centres.

The rest of the paper is structured as follows: the section 2 describes the ART network application to the object detection; the proposed ART+ network is introduced in section 3. Several experimental results of the usefulness and reliability of the approach are presented in section 4 and some conclusions are given in section 5.

2 ART-based Segmentation

In this section, a segmentation method based on ART-type networks is proposed to deal with the object detection problem. The input patterns are composed by the RGB values of the pixel. Nevertheless, other colour spaces like Lab and HSV can be used instead of RGB [7]. ART networks require no preliminary training; the first input pattern becomes the centre of cluster 1. This centre adapts on-line as successive similar patterns which are assigned to cluster 1, and if an input pattern differs sufficiently from the cluster 1 centre, it becomes the initial centre of cluster 2; and so on.

Assume that a video sequence is defined as a sequence of frames $\langle f(1), f(2), ...f(L)\rangle$ and each frame is defined as $f(k) = (x_1(k), x_2(k), ...x_{NM}(k))$, where L is the number of frames, N and M are the width and height of the images, respectively, and $x_i(k) = (r, g, b)$ is the i-th pixel of the k-th frame. The input data for the i-th ART network is defined as the following sequence $\langle x_i(1), x_i(2), ...x_i(L)\rangle$.

Note that an ART network is used for each pixel in the frame; therefore, there are as many ART networks as pixels in the scene. This kind of methods belongs to the so called pixel-level methods. The goal of each ART network is to *classify* the input patterns (pixels) as foreground or background along the time (frames).

Taking for granted that the background is more frequent than foreground in a video sequence, the most activated neurons are used to model the background. Let $a_i^T(k) = a_i^{T-1}(k-1) + \delta_i(k)$ be the activity of the neuron i at the k-th frame for the last T frames, where $\delta_i(k)$ is defined as follows

$$\delta_i(k) = \begin{cases} 1 & \text{neuron } i \text{ is the winner at } k\text{-th frame} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Note that the value of the activity of a neuron i is ranged ($a_i^T \in [0, T]$). The criterion of selecting the winner is based on a minimum distance measure (e.g., Euclidean or other measure) between the input pattern and the weight vector of the neuron. The set of background neurons B is defined as follows

$$B = \left\{ i \quad \middle| \quad \frac{1}{T}\sum_{j\in B} a_j > \theta \quad \wedge \quad \nexists C \left(|C| < |B| \quad \wedge \quad \frac{1}{T}\sum_{j\in C} a_j > \theta \right) \right\} \tag{2}$$

where $\theta \in [0, 1]$ is a specified threshold, which determine the percent of background.

3 Model Approach

This section describes the proposed ART-type network (ART+) which combines several improvements to outperform the traditional ART network for object detection. If each input pattern is viewed as a 3-dimensional vector (R,G,B) corresponding to the pixel, then the best comparative weight vector, which is denoted as a winner, can be easily obtained by a minimum distance measure. Additionally, the vigilance test is designed to check the similarity of the winner with the input pattern. If the winner passes the vigilance test, then the weight vector of the winner is adjusted according to expressions (3); otherwise, a new neuron with the input pattern is created as initial weight vector.

$$w_i(k+1) = w_i(k) + \lambda_i \|x(k) - w_i(k)\| \qquad \lambda_i = \frac{1}{1 + a_i^T} \tag{3}$$

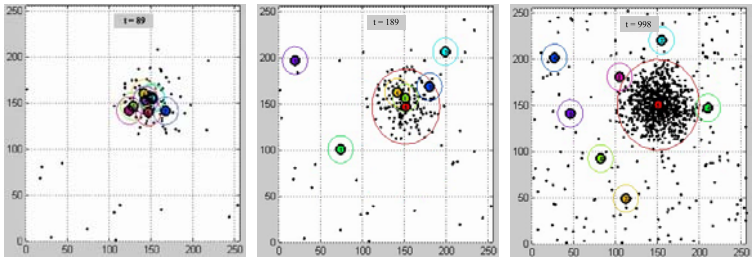
The distribution of the background and foreground data in a scene is very different (e.g. see figure 1). Background data is usually concentrated in a few distributions, i.e. background data can be typically modelled by two or three clusters. However, foreground data is more scattered than background data. Figure 1(c) shows that the dispersion of the foreground data is greater than the background data, which is modelled by only one neuron (cluster). In this sense, learning rates and vigilance parameters are introduced as neural parameters to deal with the different variability of the input data, i.e. a learning rate (λ_i) and a vigilance parameter (ρ_i) are associated to each neuron i. Note that the learning rate is initially maximum ($\lambda_i = 1$) and decreases to a minimum value ($\lambda = 1/(1 + a_i^T)$).

To reduce the number of neurons and to avoid the problem of the neurons death, clusters whose neural activity is null ($a = 0$) are removed in the ART+ network. Additionally, a neural merging process is introduced to avoid the continuously increment of neurons when the vigilance test is failed. Two close clusters can be modelled using a single larger cluster in some situations to reduce the number of neurons. Figure 1 shows how several clusters at the beginning of the scene (figure 1(a)) were merged at the end (figure 1(c)). The neural merging process consists of creating a new neuron (cluster) with a bigger vigilance parameter to cover the input data assigned to the source neurons, which are removed. Both neurons must be located very close, i.e. both clusters must be significantly overlapped, satisfying the following condition

$$\frac{\|w_i - w_j\|}{\rho_i + \rho_j} < \tau \tag{4}$$

Therefore, an overlap rate (τ) is introduced to control the overlapped grade between two clusters. This value determine the necessary overlapped grade of two clusters to be merged. Note that lower values of the overlapped rate involve in greater overlap between clusters, consequently two clusters are totally overlapped when the overlap rate is null ($\tau = 0$). The new neuron q (cluster) is created as following

$$w_q = \frac{w_i a_i + w_j a_j}{a_i + a_j} \qquad \rho_q = \frac{\rho_i a_i + \rho_j a_j}{a_i + a_j} + min\{\|w_q - w_i\|, \|w_q - w_j\|\} \tag{5}$$



(a) (b) (c)

Fig. 1: Evolution of the proposed ART+ network for a selected pixel of the scene. Circumferences around each neuron represent the vigilance parameter using Euclidean distance. Results at 89-th (a), 189-th (b) and 998-th (c) frames are showed.

An amount related to the Euclidean distance between the source neurons and the final one is added to the second expression in order to increase the vigilance threshold slightly to cover the associated patterns with the previous neurons.

4 Experimental Results

In this section a comparative study between the proposed ART+ network, the traditional ART network and other techniques mentioned in the literature is presented. Different video sequences are used to prove the usefulness and effectiveness of our approach in a variety of environments (Figure 2). The only requirement is that these sequences have the Ground Truth (GT), which is considered as the set of manually segmented images of the motion objects, to assess the results. According to the evaluation, the following performance measures are defined as

$$\text{precision} = \frac{tp}{tp + fp} \qquad \text{recall} = \frac{tp}{tp + fn} \qquad \text{F-measure} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

where tp is the number of true positives, fp is the number of false positives, and fn is the number of false negatives in the frame. F-measure combines the two prior values into a single number measure. The evaluation, which is showed in Table 1 getting the F-measure over a set of selected sequences with GT, compares the two different ART network approaches with the most popular statistical techniques based on Gaussian distributions to model the background of the scene [3, 2], which are called GMM (Gaussian Mixture Model) and PF (*Pfinder*, which uses just one single distribution) respectively. For a fair comparison, the same learning rate was used ($\lambda = 0.05$) and no post processing is applied for all the involved techniques in each test. Note that statistical values as mean and standard deviation are obtained after evaluating the output of each method in each frame of the sequence and compare it with its corresponding GT.

To measure the improvement in the ART+ network, previously it is necessary to adjust the parameters that affect the quality of the segmentation results. Setting $\tau = 0.9$ in the ART+ approach because of its insignificant impact, different configurations

Fig. 2: Experimental results of our ART+ network approach over the sequences Video2, Video4, WaterSurface and PETS2001.

Table 1: Comparison of the F-measure mean and standard deviation over the frames of three sequences of the literature.

	Video2	Video4	WaterSurface
ART	0.8503±0.024	0.338±0.042	0.5817±0.028
$ART+$	**0.8727±0.022**	*0.5335±0.016*	**0.7899±0.078**
GMM	0.6729±0.076	*0.5292±0.099*	0.3645±0.233
PF	0.6667±0.09	*0.5309±0.071*	0.5474±0.534

for both ART and ART+ approaches, combining the values of the parameters $\lambda = \{0.001, 0.005, 0.01, 0.05, 0.1\}$ and $\rho_{ini} = \{15, 20, 25\}$, are generated. Table 1 shows that the ART+ network achieves definitely better results for the sequences Video2 and WaterSurface and remains approximately the same for Video4, in comparison to the statistical approaches GMM and PF. Moreover, the results in figure 3 demonstrate that the segmentation quality of the ART+ network outperforms the traditional ART.

5 Conclusions

A new approach based on ART-type networks is developed, with the aim of modelling clustering online problems that requires discarding information and unsupervised learning. In this work, this method is applied to detect objects in motion in video sequences. The goal of the proposed ART+ network is to adapt to the pixel behaviour along the time, by clustering the input data as foreground or background.

Due to unfavourable factors, such as illumination variance, shadows and shaking branches, the background segmentation is a difficult problem to be treated with the traditional ART, since the variability of the input data is high and diverse among the pixels of a scene. The results show that the introduced neural learning rate and neural vigilance threshold improve the quality of the segmentation. Moreover, a neural merging process is also proposed to obtain a more effective and robust representation of the input data,

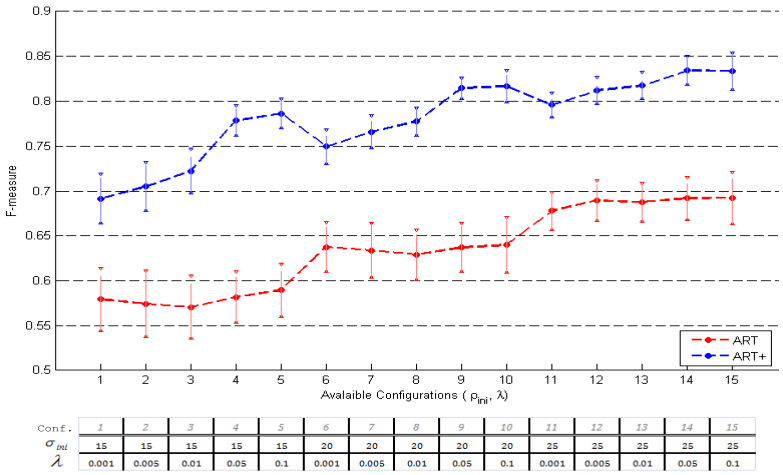


Fig. 3: F-measure comparison in the 'WaterSurface' sequence between the two network approaches ART and ART+, using different parameter settings of λ and ρ_{ini}.

and to reduce the computational time of the proposed ART+ network. Additionally, a removing inactive neurons mechanism is performed to treat the traditional problem of dead neurons.

Acknowledgements

This work is partially supported by Junta de Andalucía (Spain) under contract TIC-01615, project name Intelligent Remote Sensing Systems.

References

[1] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell. Towards robust automatic traffic scene analysis in real-time. In anonymous, editor, *Proceedings of the International Conference on Pattern Recognition*, pages 126–131, 1994.

[2] C.R. Wren, A. Azarbayejani, T. Darrell, and A. Pentl. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[3] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In anonymous, editor, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 246–252, 1999.

[4] I. Haritaoglu, D. Harwood, and L.S. Davis. W4: real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000.

[5] S. Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11(1):23–63, 1987.

[6] G. Carpenter and S. Grossberg. Art-2 self-organization of stable category recognition codes for analog input patterns. *Appl. Optics*, 26(9):4919–4930, 1987.

[7] C. Benedek and T. Sziranyi. Study on color space selection for detecting cast shadows in video surveillance. *International Journal of Imaging Systems and Technology*, 17(3):190–201, 2007.