A Dipolar Competitive Neural Network for Video Segmentation

R.M. Luque[1], D. López-Rodríguez[2], E. Dominguez[1], and E.J. Palomo[1]

[1] Department of Computer Science, University of Málaga, Málaga, Spain
{rmluque,enriqued,ejpalomo}@lcc.uma.es
[2] Department of Applied Mathematics, University of Málaga, Málaga, Spain
dlopez@ctima.uma.es

Abstract. This paper present a video segmentation method which separate pixels corresponding to foreground from those corresponding to background. The proposed background model consists of a competitive neural network based on dipoles, which is used to classify the pixels as background or foreground. Using this kind of neural networks permits an easy hardware implementation to achieve a real time processing with good results. The dipolar representation is designed to deal with the problem of estimating the directionality of data. Experimental results are provided by using the standard PETS dataset and compared with the mixture of Gaussians and background subtraction methods.

1 Introduction

The aim of moving object segmentation is to separate pixels corresponding to foreground from those corresponding to background. This task is complex by the increasing resolution of video sequences, continuing advances in the video capture and transmission technology.

The process of background modeling by comparison with the frames of the sequence is often referred to as background subtraction. These methods are widely exploited in videos for moving object detection. Adaptive models are typically used by averaging the images over time [1,2,3] creating a background approximation. While these method are effective in situations where objects move continuously, they are not robust in scenes with many moving objects, particularly if they move slowly.

Wren et al. [4] used a multiclass statistical representation based on Gaussian distributions, in which the background model is a single Gaussian per pixel. A modified version modeling each pixel as a mixture of Gaussians is proposed by Stauffer and Grimson [5]. This statistical approach is robust in scenes with many moving objects and lighting changes, and it is one of the techniques most cited in the literature.

Several steps are required for a typical video surveillance system to reach the objective. At first, a video object segmentation method is required to obtain the objects in motion of the stream. Subsequently, a tracking algorithm is applied to identify the objects in several frames of the sequence. A matching between each blob (or set of blobs) and an object previously recognized has to be done. Finally, an algorithm to detect the object behavior it is used to understand and analyze everything what it is happening in

H. Geffner et al. (Eds.): IBERAMIA 2008, LNAI 5290, pp. 103–112, 2008.

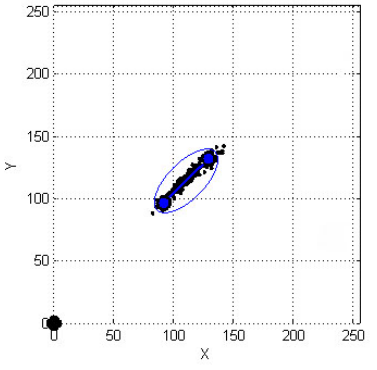


Fig. 1. A dipole is able to capture the intrinsic directionality of a set of data points

a scene. Therefore, low time complexity is required at the object segmentation stage in order to carry out the entire process in real time.

In this work an unsupervised competitive neural network is proposed for object segmentation in real time. The proposed neural approach is based on a dipolar representation in order to achieve a better representation of data since it is able to capture the intrinsic directionality of data at a low computational cost. Although the mixture of Gaussians model can obtain the directionality of data, the process of computing the covariance matrix is highly expensive from the computational point of view. Thus, in practice, it is usually assumed the independence of the RGB components, then all information relative to directionality is lost in this simplified model [6].

Dipolar competitive neural networks (DCN) [7] differ from traditional competitive networks [8,9,10] in that every prototype w_j is now represented by a segment formed of two distinct vectors $w_j^{(1)}$ and $w_j^{(2)}$. These two vectors represent the end-points of a segment in the input space. Note that $w_j^{(1)}$ and $w_j^{(2)}$ can also be interpreted as the two foci of an ellipsoid. The classical competitive learning rule can be applied to this model to make this segment adjust to data, obtaining the directionality of patterns activating that dipole (that is, patterns whose nearest dipole is the given one), see Fig. 1.

2 A Neural Model for Color Video Segmentation

In this work, a classification task is locally performed, for each pixel in the video sequence, in parallel. The classification algorithm used is a competitive network based on dipoles.

Let us consider a DCN for each pixel in the video. Each of these networks models the RGB space associated to the corresponding pixel, that is, training patterns are formed by the three values (R,G,B) for the pixel.

The target of the network is to classify the input pattern (for the specified pixel) at each frame as foreground or background. It can be noted that our model allows the use of many neurons (also called dipoles in this particular case) to represent multimodal classes. In our case, three neurons have been used to model the scene including both

background and foreground. The B most activated neurons are used to model the background, whereas the rest of neurons correspond to foreground objects. This value B is computed as the amount of neurons whose number of activations $n_{a_1}, \ldots, n_{a_B}$ verify $\frac{n_{a_1}+\ldots+n_{a_B}}{N} > T$ for a prefixed threshold T, where N is the total number of activations of all neurons, as proposed in [6]. In this work, we have used $T = 0.7$.

The use of dipoles has an additional advantage for the classification task: the directionality of data is learned, thus obtaining extra information about the shape of clusters.

A detailed description of this model is presented in the next two subsections. There, we study the activation of the winning dipole by defining an adequate synaptic potential for each dipole, and the learning rule used to update each of the foci of the ellipsoids.

2.1 Definition of Synaptic Potentials

The computation of the synaptic potential received by each dipole is based on a crisp-fuzzy hybrid neighborhood, which enables to define certain mechanisms that improve the performance of the neural model.

For each neuron, there is a value r_j, defining the crisp neighborhood of the corresponding prototype $w_j = (w_j^{(1)}, w_j^{(2)})$: $\mathcal{N}_j = \{x : \|x - w_j^{(1)}\| + \|x - w_j^{(2)}\| \leq 2r_j\}$. This definition corresponds to an ellipsoid whose foci are $w_j^{(1)}$ and $w_j^{(2)}$ and such that the length of the main semi-axis is r_j. Note that the main direction of the ellipsoid corresponds to the direction of the dipole $w_j = (w_j^{(1)}, w_j^{(2)})$.

The fuzzy neighborhood of neuron j is given by a membership function μ_j defined over points not belonging to $\mathcal{N}_j$, and taking values in the interval $(0, 1)$. Usually, the membership function present in this model is of the form:

$$\mu_j(x) = \mathrm{e}^{-k_j\left(\|x-w_j^{(1)}\|+\|x-w_j^{(2)}\|-2r_j\right)}, \quad \text{for } x \notin \mathcal{N}_j \tag{1}$$

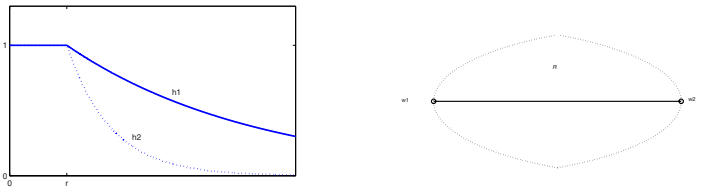
The items above provide us with a reasonable way to select the winning dipole for each input pattern, which is the one to be updated at the current iteration, after the input pattern is received by the network. We define the synaptic potential received by dipole j when pattern x is presented to the network as $h_j(x) = 1$ if $x \in \mathcal{N}_j$ and $h_j(x) = \mu_j(x)$ otherwise.

The winning dipole (with index denoted as $q(x)$) is the one receiving the maximum synaptic potential: $h_{q(x)} = \max_j h_j(x)$.

To break ties (more common when two crisp neighborhoods, corresponding to different dipoles, are overlapped), we consider the dipole which has been activated more times as the winning dipole. That is, if, for every j, n_j counts the number of times dipole j has been winner, and pattern x belongs to $\mathcal{N}_{j_1} \cap \mathcal{N}_{j_2}$ (overlapped neighborhoods), then $q(x)$ is defined as k for which $n_k = \max\{n_{j_1}, n_{j_2}\}$. In case $n_{j_1} = n_{j_2}$, a random dipole in $\{j_1, j_2\}$ is selected as winner.

The use of this hybrid neighborhood allows us to better assign an input pattern to a class:

- If input pattern x belongs to $\mathcal{N}_j$, the network assumes that the best matching class for x is the associated to dipole w_j.



(a) (b)

Fig. 2. (a) Comparison between the synaptic potentials h_1 and h_2 of two dipoles, such that $k_2 > k_1$, supposing $r_1 = r_2 = r$. (b) R represents the area of the input space well-represented by the dipole.

– If input pattern x does not belong to any crisp neighborhood, its most likely class is represented by the dipole achieving the maximum value of the membership function for this particular input.

The value of the parameter k_j is related to the slope of the function. The higher its value, the higher the slope is. For a great value of k_j, the fuzzy neighborhood of w_j will be more concentrated around $\mathcal{N}_j$. The effect (on the corresponding h_j) of increasing the value of k_j is shown in Fig. 2(a).

With the help of this parameter, we can model a mechanism of consciousness, necessary to avoid dead neurons. When a neuron is activated and updated many times, its crisp neighborhood usually englobes a very high percentage of the patterns associated to the neuron. In this case, patterns outside this neighborhood are not likely to belong to the corresponding category. Thus, the membership function of the fuzzy neighborhood should be sharp, with a high slope.

If a dipole, j, is rarely activated, then the ellipsoid $\mathcal{N}_j$ does not represent accurately the patterns associated to the dipole. Thus, the membership associated to the fuzzy neighborhood should express the actual fuzziness present in the data and therefore assign higher values to patterns outside $\mathcal{N}_j$.

From all this we can deduce that a good way to define k_j is proportional to the number of times that neuron j has been activated, n_j.

2.2 The Learning Rule

In what follows, let us suppose that the winning dipole is $w = (w^{(1)}, w^{(2)})$.

The purpose of the learning rule described in this section is to make each dipole, and the associated ellipsoid, represent both the location and distortion (with respect to the centroid of the dipole) of its corresponding data as accurately as desired.

Let us denote $R = \{x : \|x - w^{(1)}\|, \|x - w^{(2)}\| < \|w^{(1)} - w^{(2)}\|\}$. R denotes the set of points whose distance to each focus is lower than the distance between foci. A reasonable criterion to determine whether a given data point x is well-represented by the ellipsoid, is that $x \in R$, see Fig. 2(b) for a graphical representation of this situation. In this case, the dipole has just to be adjusted to better represent data location, trying to minimize the distance from x to the dipole centroid $\overline{w} = \frac{w^{(1)}+w^{(2)}}{2}$, $\|x - \overline{w}\|^2$.

The updating scheme in this case, at iteration step t, is therefore:

$$w^{(i)}(t+1) = w^{(i)}(t) + \lambda(x - \overline{w}(t))$$

for $i \in \{1, 2\}$, λ being the learning rate parameter. This means that the centroid is updated as follows: $\overline{w}(t+1) = \frac{w^{(1)}(t+1)+w^{(2)}(t+1)}{2} = \overline{w}(t) + \lambda(x - \overline{w}(t))$ that is, the ellipsoid is able to better capture the location of x.

When $x \notin R$, the situation changes. It does not seem reasonable to update the dipole as in the previous case. In this paper, we propose an updating scheme based on how well points activating the dipole are represented by the latter.

To this end, let us define n_{in} as the number of points activating the dipole that belong to R at current iteration. Analogously, n_{out} is the number of points activating the dipole which do not belong to R.

A measure of how well points are represented by the dipole is given by the quotient $\frac{n_{\text{out}}}{n_{\text{in}}} = \rho$. Given $c \in (0, 1]$, we say that the dipole represents accurately the data points activating it if, at least, a $100c\%$ of these points belongs to R, that is, c is the fraction of points activating the dipole and belonging to R. This means that $100(1-c)\%$ of the points does not belong to R, so $\rho \leq \frac{1-c}{c}$. Let us denote $\rho_0 = \frac{1-c}{c}$ the maximum desired value for ρ. Note that c is an user-defined parameter.

Then, there are two cases if $x \notin R$:

- If $\rho > \rho_0$, then there are many data points outside R comparing with the number of points that belong to R, so the dipole does not represent well the dataset. In order to improve this representation, the two foci will be updated towards the input data x, as in the standard competitive learning rule:

  $$w^{(i)}(t+1) = w^{(i)}(t) + \lambda(x - w^{(i)}(t))$$

  for $i \in \{1, 2\}$.
- If $\rho \leq \rho_0$, then points are very well-represent by the ellipsoid, according to our definition. Thus, it suffices to update the focus nearest to x, denoted by $w^{(s(x))}$:

  $$w^{(s(x))}(t+1) = w^{(s(x))}(t) + \lambda(x - w^{(s(x))}(t))$$

  This transition allows to capture the actual directionality of data associated to the dipole.

This complex updating scheme helps the network to reduce the dispersion of the data points with respect to the dipole centroid.

On the other hand, our main concern is to detect the main directionality of data which is given by the main direction of the dipoles. Empirical studies have revealed an underlying ellipsoidal structure of data in the RGB color space, with a clearly predominant main direction. Since non-principal directions are not of importance in this problem we consider the corresponding semi-axes fixed. Therefore, only r_j, the major semi-axis, is updated accordingly in each step.

3 Improving the Segmentation

In this work, we have studied two improvements of the segmentation process:

Spurious Object Detection. In most real-life situations, some processes, such as compression, decrease the sharpness of the video sequence. For this reason, many objects consisting in a single and separated pixel are detected by segmentation methods (spurious objects). To solve this issue, we propose the use of a post-processing method, which consists in finding those isolated pixels representing objects and marking them as background. After this process, for each of those pixels, the corresponding dipole representing background is updated (by using the proposed learning rule), whereas the dipole which had been updated (representing foreground), returns to its state previous to the processing of the current frame.

Shadow Detection. Objects in motion probably cast shadow on the background, confusing with foreground pixels and interfering the correct detection of the scene objects. In our system, we develop the technique proposed in [11], based on the proportionality property between the shadow and the background in the RGB color space.

With these enhancements, we obtain better segmentation results, as will be shown in next section.

4 Results

In this section a comparison between our proposed neural approach and other techniques mentioned in the literature is done. We use different video sequences obtained from Internet to demonstrate the effectiveness of our algorithm for background subtraction and foreground analysis in a variety of environments. These sequences also present different features, from diverse kind of lighting to distinct objects in motion (people, vehicles) in order to conduct a more comprehensive study of the proposed method. Note that same parameters were used for all scenes.

Figure 3 shows the results obtained after applying the studied techniques. Our DCN model is compared with the mixture of Gaussians model (MoG) [6] and with another typical algorithm of background subtraction [3], consisting of subtracting the processed frame from a background model previously computed. A fixed threshold value has been established to get the objects segmented. In MoG we have set the number of normal distributions to $K = 3$, the mean of each distribution is initially computed as the gray-level value of the corresponding pixel in the first frame of the sequence, and the standard deviation is initialized to 25.

In our neural approach, fine-tuning of the model initial parameters (number of dipoles, learning rates, initial major semi-axis r_j) is made before the segmentation

Table 1. Comparative analysis of the success rate among the studied methods for the sequence observed in Fig. 4

Method	% Matching	% FP	% FN
DCN+	99.6957	0.07068	3.408
DCN	99.4387	0.2895	4.8602
MoG	98.7639	1.0967	1.3939
BS	98.6636	0.28041	30.1686

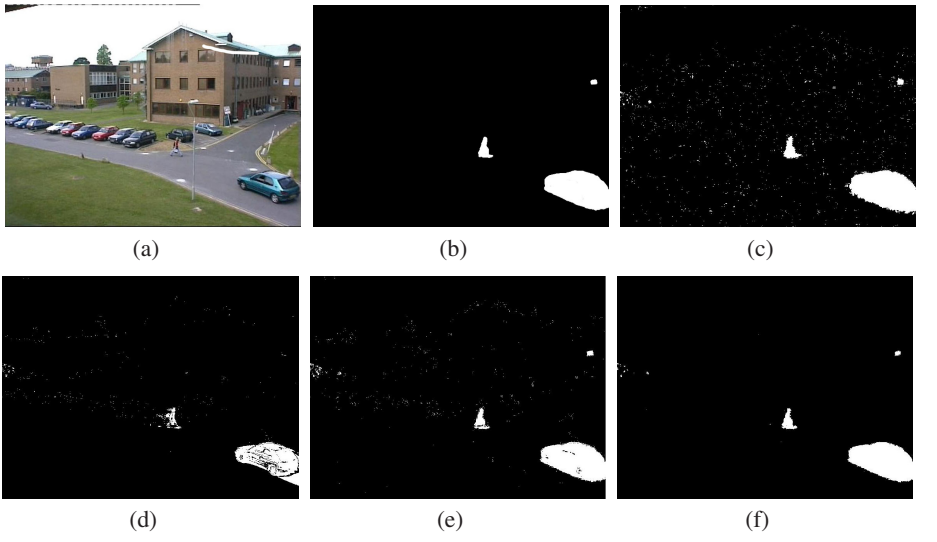
(a) (b) (c) (d)

Fig. 3. Results of applying three segmentation methods to several frames. (a) column shows the capture frames for each scene in raw form; (b) proposed method; (c) mixture of Gaussians method; (d) background subtraction.

process. Three dipoles have been used to model the scene including both background and foreground, although more dipoles can be added for a better representation of multimodal backgrounds. The learning rate is decreased after each frame until stabilizing at a fixed value, while the major semi-axis is initialized to 15. Both the simple model with neighborhoods (DCN) and the extended model (DCN+), in which a mechanism to correcting the model and to avoid spurious pixels is applied, have been tested.

Figure 3 shows an example result on one PETS 2001 (IEEE Performance Evaluation of Tracking and Surveillance Workshops) sequence. A quantitative comparison among the different algorithms is shown in Tables 1 and 2. Three measures have been defined to evaluate the performance of each method. A false positive rate (FP) shows the number of wrong background pixels and a false negative rate (FN) is used to show the number of misleading foreground pixels. The success rate, indicating the accuracy of the corresponding algorithm, is presented in the column labelled '% Matching'. By observing these results, it can be noted that in general, our method gets better segmentation than the rest of analyzed methods, when we compare with a ground truth image. It is remarkable the ability of our neural approach to efficiently adapt to diverse scenes, without any modification of the mentioned parameters. Another comparison can be showed in Fig. 5 by using the intelligent room sequence, available at http://cvrr.ucsd.edu/aton/shadow.

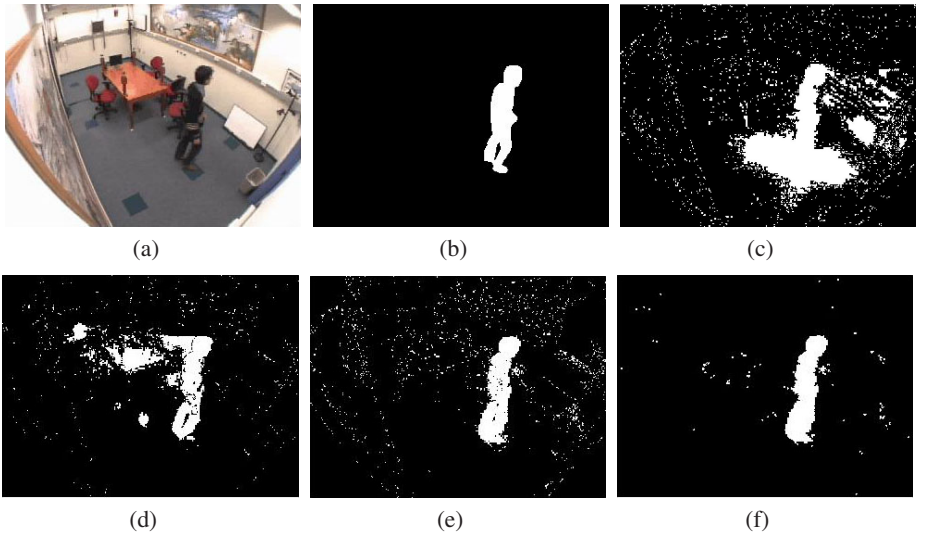
Figure 6 shows the final result of our approach, after applying a shadow detection method and enhancing the obtained frame removing spurious objects. As we can observe, this segmentation is effective enough to be used in a subsequent tracking phase.



(a) (b) (c)

(d) (e) (f)

Fig. 4. Comparison among the studied techniques: (a) a frame obtained from the PETS01 sequence in raw form; (b) ground truth; (c) mixture of gaussians method; (d) background subtraction method; (e) DCN; (f) DCN+



(a) (b) (c)

(d) (e) (f)

Fig. 5. Comparison among the studied techniques: (a) a frame of the intelligent room sequence; (b) ground truth; (c) mixture of gaussians method; (d) background subtraction method; (e) DCN; (f) DCN+

Table 2. Comparative analysis of the success rate among the studied methods for the sequence observed in Fig. 5

Method	% Matching	% FP	% FN
DCN+	98.9388	0.86961	6.7224
DCN	97.2409	2.5846	7.9157
MoG	87.9193	12.0628	12.6094
BS	95.4362	4.1031	18.1782

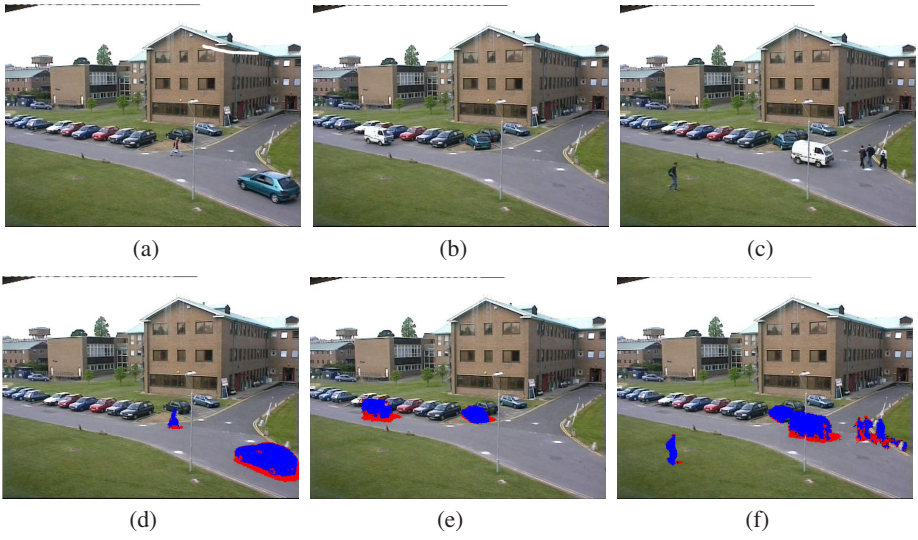


Fig. 6. Final results of the proposed method. In (a), (b) and (c) we can observe three frames (500, 750, 867) of a scene from PETS01 in raw form. (d), (e) and (f) show the segmentation results using our neural approach (DCN+), after applying shadow detection.

5 Conclusions and Future Work

In this work a new competitive neural network based on dipoles for video object detection and segmentation is presented. An unsupervised learning is performed to model the RGB space of each pixel together with the directionality of data using a dipolar representation.

The idea of using dipoles instead of a single point in the RGB color space permits to obtain the direction of the input pixel data. In this sense, experimental results have shown that the background model composed of ellipsoidal shapes (represented by the dipoles) outperforms the accuracy of other methods.

The segmentation accuracy of the proposed neural network is compared to mixture of Gaussian (MoG) and background subtraction (BS) models. In all the performed comparisons, our model achieved better results in terms of success rate and false positive rate, whereas the false negative rate is, at least, comparable to the obtained by the other studied methods.

Moreover, the proposed algorithm can be parallelized on a pixel level and designed to enable efficient hardware implementation to achieve real-time processing at great frame rates.

Other applications of this model will be studied, including the incorporation of this neural network to a remote sensing system performing people and vehicles tracking on closed scenes.

Acknowledgements

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

References

1. Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: Detecting moving objects, ghosts, and shadows in video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(10), 1337–1342 (2003)
2. Koller, D., Weber, J., Huang, T., Malik, J., Ogasawara, G., Rao, B., Russell, S.: Towards robust automatic traffic scene analysis in real-time. In: Proceedings of the International Conference on Pattern Recognition (1994)
3. Lo, B., Velastin, S.: Automatic congestion detection system for underground platforms. In: Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 158–161 (2001)
4. Wren, C., Azarbayejani, A., Darrell, T., Pentl, A.: Pfinder: Real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 780–785 (1997)
5. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1999)
6. Stauffer, C., Grimson, W.: Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 747–757 (2000)
7. García-Bernal, M., Muñoz, J., Gómez-Ruiz, J., Ladrón De Guevara-Lopez, I.: A competitive neural network based on dipoles. In: Mira, J., Álvarez, J.R. (eds.) IWANN 2003. LNCS, vol. 2686. Springer, Heidelberg (2003)
8. Oliveira, P., Romero, R.: Improvements on ica mixture models for image pre-processing and segmentation. Neurocomputing (in press, 2008)
9. Mu-Chun, S., Hung, C.H.: A neural-network-based approach to detecting rectangular objects. Neurocomputing 71(1-3), 270–283 (2007)
10. Meyer-Base, A., Thummler, V.: Local and global stability analysis of an unsupervised competitive neural network. IEEE Transactions on Neural Networks 19(2), 346–351 (2008)
11. Horprasert, T., Harwood, D., Davis, L.S.: A statistical approach for real-time robust background subtraction and shadow detection. In: Proceedings of International Conference on Computer Vision (1999)