A Neighborhood-Based Competitive Network for Video Segmentation and Object Detection

R.M. Luque Baena[1], E. Dominguez[1], D. López-Rodríguez[2], 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 work proposes an unsupervised competitive neural network based on adaptive neighborhoods for video segmentation and object detection. The designed neural network is proposed to form a background model based on subtraction approach. The synaptic weights and the adaptive neighborhood of the neurons serve as a model of the background and are updated to reflect the statistics of the background. The segmentation performance of the proposed neural network is examined and compared to mixture of Gaussian models. The proposed algorithm is parallelized on a pixel level and designed to enable efficient hardware implementation to achieve real-time processing at great frame rates.

1 Introduction

The goal of the video 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. As a result, research into more efficient algorithms for real-time object segmentation continues unabated.

The process of modeling the background by comparison with the frames of the sequence is often referred to as background subtraction. These methods are widely exploited for moving object detection in videos. Detection of moving is the next step of information extraction in many computer vision applications, such as video surveillance and people tracking. In these applications, robust tracking of objects is required for a reliable and effective moving object detection. While the fast execution and flexibility in different scenarios (indoor, outdoor) or different light conditions should be considered basic requirements to be met, precision is another important goal. In fact, a precise moving object detection makes tracking more reliable and faster.

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 to scenes with many moving objects particularly if they move slowly. Toyama et al. [4] propose the Wallflower algorithm in which background maintenance and background subtraction are carried out at pixel, region, and frame levels. Haritaoglu et al. [5] build a statistical model by representing each pixel with three values: its minimum and maximum intensity, and the maximum intensity difference between consecutive frames, which are updated periodically.

V. Kůrková et al. (Eds.): ICANN 2008, Part I, LNCS 5163, pp. 877–886, 2008.

McKenna et al. [6] use an adaptive background model with color and gradient information to reduce the influences of shadows and unreliable color cues.

Wren et al. [7] used a multiclass statistical model based on Gaussian distributions. But 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 [8]. This approach is robust to scenes with many moving objects and lighting changes, but it is only able to achieve real time processing of small video formats (120x160 pixels).

Neuronal networks have been widely used for images segmentation. There exists numerous works in the fields of typewriting recognition [9] and medical images segmentation [10,11], but many of these neural techniques are not suitable for real time applications by their long time of computation in the training process. The development of a parallelized object segmentation approach, which would allow for object detection in real time for complex video sequences, is the focus of this paper. Neural networks posse intrinsic parallelism which can be exploited in a suitable hardware implementation to achieve fast segmentation of foreground objects.

In this work an unsupervised competitive neural network is proposed for objects segmentation in real time. The proposed approach employs is based on competitive neural network to achieve background subtraction. A new unsupervised competitive neural network based on neighborhood is designed to serve both as an adaptive model of the background in a video sequence and a classifier of pixels as background or foreground. The segmentation performance of the proposed neural network is qualitatively examined and compared to mixture of Gaussian models.

2 A Neural Model for Color Video Segmentation

Let us consider a competitive network for each pixel in the video. Each of these networks models the RGB space associated to the corresponding pixel. Thus, every pattern is formed by the three values (R,G,B) for the pixel. Then, the sequence of input patterns learned by each network consists on the different (R,G,B) values of the same pixel in every video frame.

The target of the network is to classify the input pattern (for the specified pixel) at each frame as foreground or background. Each neuron is used to model only one class. It can be noted that our model allows the use of many neurons to represent multimodal classes.

Our model performs a clustering based on a crisp-fuzzy hybrid neighborhood. For each neuron, there is a value r_j, representing the crisp neighborhood of the corresponding synaptic weight w_j: $\mathcal{N}_j = \{x : \|x - w_j\| \leq r_j\}$. The fuzzy neighborhood of neuron j is given by a membership function μ_j defined over the entire input space, and taking values in the interval $(0, 1)$.

Usually, the membership function present in this model is of the form: $\mu_j(x) = \mathrm{e}^{-k_j(\|x-w_j\|-r_j)}$

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. 1.

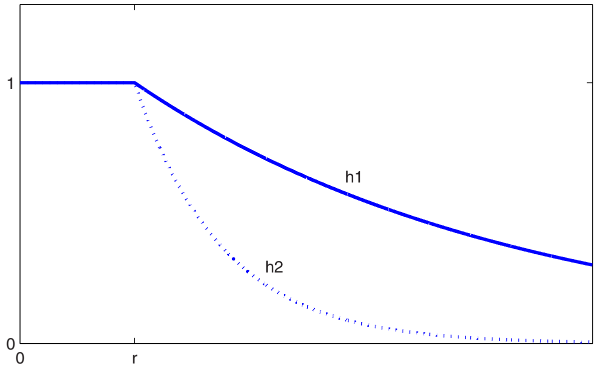

Fig. 1. Comparison between the synaptic potentials h_1 and h_2 of two neurons, such that $k_2 > k_1$, supposing $r_1 = r_2 = r$

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

Thus, a good way to define k_j is proportional to the number of times that neuron j has been activated, n_j.

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

- If input pattern x belongs to $\mathcal{N}_j$, the network assumes that the best matching category for x is the associated to w_j.
- If input pattern x does not belong to any crisp neighborhood, its most likely category is represented by the neuron achieving the maximum value of the membership function for this particular input.

The fuzzy neighborhood also allows to include other mechanisms in the learning phase, such as neuron consciousness.

The items above provide us with a reasonable way to select the winning neuron 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 neuron j when pattern x is presented to the network as

$$h_j(x) = \begin{cases} 1, & \text{if } x \in \mathcal{N}_j \\ \mu_j(x), & \text{otherwise} \end{cases} \quad (1)$$

The winning neuron (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 neurons, are overlapped), we consider the neuron which has been activated more times as the winning neuron. That is, if, for every j, n_j counts the number of times neuron 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 neuron in $\{j_1, j_2\}$ is selected as winner.

The learning rule used by each network to model the input space is the standard competitive learning rule:

$$w_{q(x)}(t+1) = w_{q(x)}(t) + \lambda \cdot (x - w_{q(x)}(t))$$

which can be viewed as the stochastic gradient descent technique to minimize the squared error function (also named distortion function):

$$F(W) = \sum_i \|x_i - w_{q(x_i)}\|^2 \tag{2}$$

where W represents a matrix whose rows are the w_j and λ is the so-called learning rate parameter, usually decreasing to 0.

However, an undesirable effect takes place under some conditions: the overlapping of two or more neighborhoods. Our model can be extended to solve this problem. We can consider a new formulation in order to avoid such overlaps.

Let us denote $Q(x) = \{j \neq q(x) : x \in \mathcal{N}_j\}$ the set of the indices j such that $x \in \mathcal{N}_j \cap \mathcal{N}_{q(x)}$, i.e., neuron j and $q(x)$ overlap at x.

In this case, neurons j with $j \in Q(x)$ may be repelled from the data, getting away from the winning neuron (which gets closer to the data) and reducing the overlap. Therefore, one should maximize the quantity $\|x - w_j\|^2$ with $j \in Q(x)$.

On the other hand, an overlap indicates that the crisp neighborhood of the winning neuron is not big enough to englobe all its associated patterns. Thus, a way to solve this problem is to increase the value of r_j, that is, to maximize (to some extent) the value of the radius r_j^2.

Let us find a learning rule with these properties, starting from an objective function, similar to Eq. (2), to be minimized.

Note that there is overlap if, and only if, $Q(x)$ is not empty. Define $\delta_{Q(x)} = 1$ if $Q(x) \neq \varnothing$, otherwise $\delta_{Q(x)} = 0$.

Then, the objective function to be optimized can be expressed in the following terms:

$$F(W, r) = \sum_i \left[\|x_i - w_{q(x_i)}\|^2 - \sum_{j \in Q(x_i)} \|x_i - w_{q'(x_i)}\|^2 - \delta_{Q(x_i)} r^2_{q(x_i)} \right] \tag{3}$$

The minus sign expresses the fact that the second and third terms must be maximized. It can be observed that the repulsion of neurons $j \in Q(x_i)$, and the increase of the radius associated to the winner, only take place when an overlap occurs.

Note that, in this new formulation, the value r_j of the radius of $\mathcal{N}_j$ is treated as an adaptive parameter, and is automatically recalculated after an overlap is detected.

The learning rule associated to the new model can be obtained by applying the stochastic gradient descent technique to optimize the function given in Eq.(3). This technique has been used in other works [12] to develop new competitive learning rules from a family of objective functions.

To apply this technique, let us consider a single term of the objective function $F(W, r)$ (denote $x = x_i$, for simplicity):

$$T(W, r) = \frac{1}{2}\left(\|x - w_{q(x)}\|^2 - \sum_{j \in Q(x)} \|x - w_{q'(x)}\|^2 - \delta_{Q(x)} r^2_{q(x)} \right)$$

Then, the learning rule is derived by differentiating the above expression with respect to w_j and r_j, and making $w_j(t+1) = w_j(t) - \lambda \frac{\partial T}{\partial w_j}$ and $r_j(t+1) = r_j(t) - \lambda \frac{\partial T}{\partial r_j}$.

$$\frac{\partial T}{\partial w_j} = \begin{cases} w_j - x, \text{ if } j = q(x) \\ x - w_j, \text{ if } j \in Q(x) \\ 0, \qquad \text{otherwise} \end{cases}$$

$$\frac{\partial T}{\partial r_j} = \begin{cases} -r_j, \text{ if } (j = q(x)) \wedge (\delta_{Q(x)} = 1) \\ 0, \qquad \text{otherwise} \end{cases}$$

Thus, we obtain the learning rule:

$$w_{q(x)}(t+1) = w_{q(x)}(t) + \lambda(x - w_{q(x)}(t)) \tag{4}$$

$$w_j(t+1) = w_j(t) - \lambda(x - w_j(t)) \quad \text{if } j \in Q(x) \tag{5}$$

$$r_{q(x)}(t+1) = (1+\lambda) r_{q(x)}(t) \quad \text{if } \delta_{Q(x)} = 1 \tag{6}$$

This constructive way of deducing this learning rule (by using the stochastic gradient descent method) ensures the minimization of the objective function of our model, given in Eq. (3).

3 Results

In this section the performance of the proposed neural approach is evaluated by testing our system on several image sequences. This video sequences are associated to different types of scenes obtained by means of a web camera or downloaded from internet. We have used very heterogeneous scenes, with different sizes and compressed by different techniques. These sequences also presented 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. Some of the scenes have been used and referenced in the literature.[1]

The most remarkable aspects evaluated in this paper, which decide whether our proposal may be used in a real-world environment, are reliability and robustness. Therefore, our system has been compared with other standard algorithms designed to perform

[1] Some of them are available at http://cvrr.ucsd.edu/aton/shadow

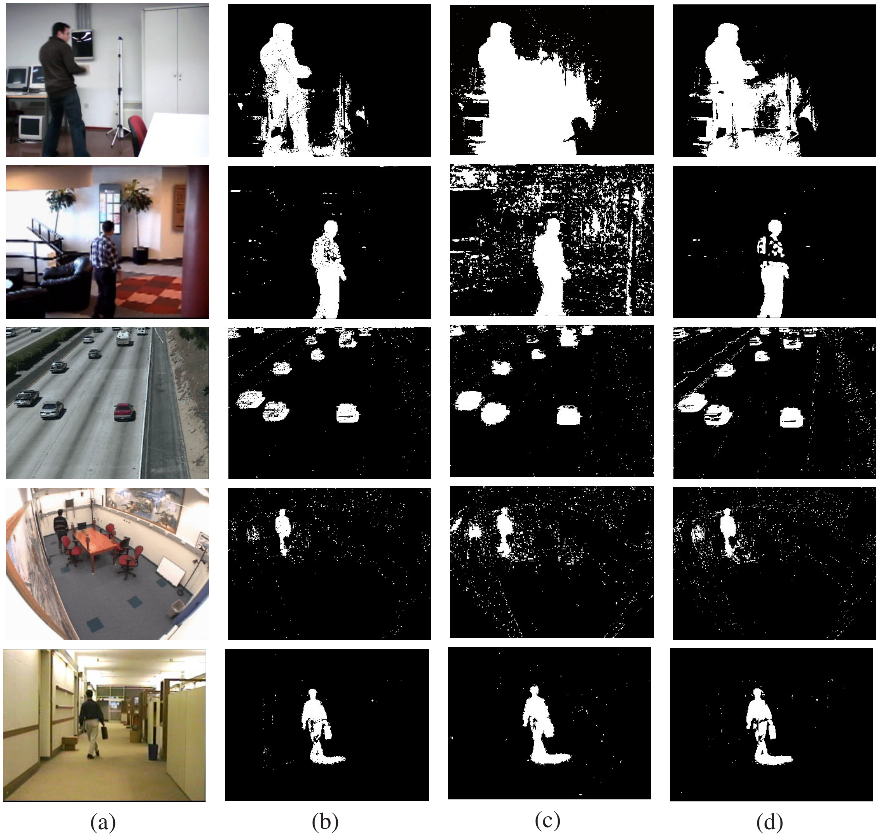

Fig. 2. Results of applying three segmentation methods to several frames. (a) column shows the capture frames for each scene in raw form; (b) proposed method; (c) mixture of Gaussians method; (d) background subtraction.

object segmentation in video sequences. Concretely, we compare with one of the most cited techniques in the specialized literature, the mixture of Gaussians model (MoG) [13]. This statistical model is based on a mixture of Gaussian distributions, able to solve correctly sudden illumination changes, multimodal backgrounds and deal with both indoor and outdoor scenes. In this work, we have set the number of normal distributions to $K = 3$. The mean of each of these distributions is initially computed as the gray-level value of the corresponding pixel in the first frame of the sequence. The other parameter, the standard deviation, is initialized to 25. Additionally, the results of another simpler technique (Background Subtraction, BS) have also been analyzed. This algorithm [3] consists on subtracting the processed frame from a background model previously computed. A fixed threshold value has been established for the whole set of analyzed scenes.

In our neural approach, fine-tuning of the model initial parameters (number of neurons, learning rate, initial radius r_j) is made before the segmentation process. These

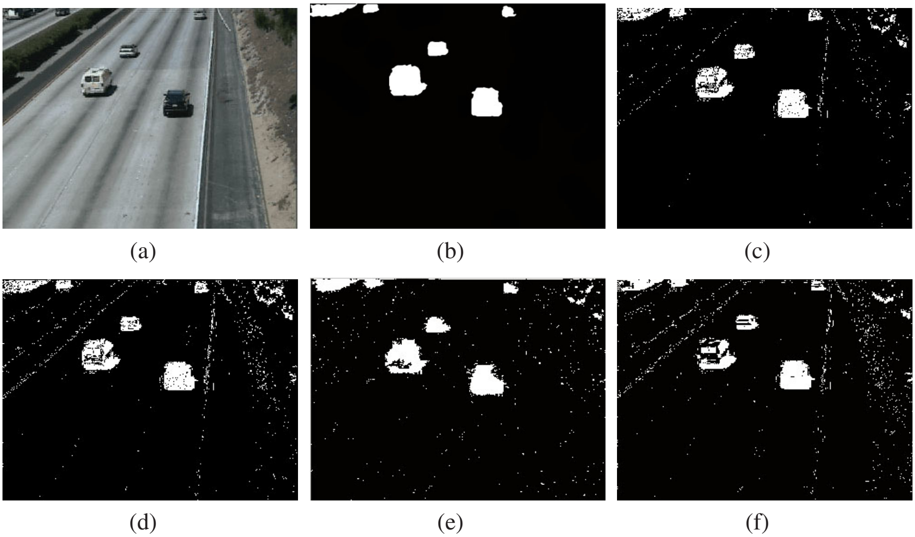

Fig. 3. Comparison among the studied techniques: (a) a traffic frame obtained in raw form; (b) ground truth; (c) neural network approach with repulsion; (d) neural network approach without repulsion; (e) mixture of gaussians method; (f) background subtraction method

Table 1. Comparative analysis of the success rate among the studied methods for the sequence observed in Fig. 3

Method	% Matching	% FP	% FN
NCM+	98.255	0.750	23.822
NCM	97.713	1.427	21.347
MoG	97.336	1.842	20.894
BS	97.125	1.772	27.355

parameters remain the same for all the scenes analyzed. Three neurons have been used to model the scene including both background and foreground, although more neurons can be added for a better representation of multimodal backgrounds. 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 [13]. In this work, we have used $T = 0.7$. The learning rate is decreased after each frame until stabilizing at a fixed value, while the radius of the neural neighborhood is initialized to 25. Both the simple model with neighborhoods (NCM) and the extended model (NCM+), in which a repulsion mechanism is used to avoid the overlapping of two neighborhoods (see Eqs. (4)-(6)), have been tested.

Figure 2 shows comparison results among different techniques for several video sequences. It can be noted a better silhouette detection of the objects using our neural approach, removing also great amount of noise regarding to the others algorithms.

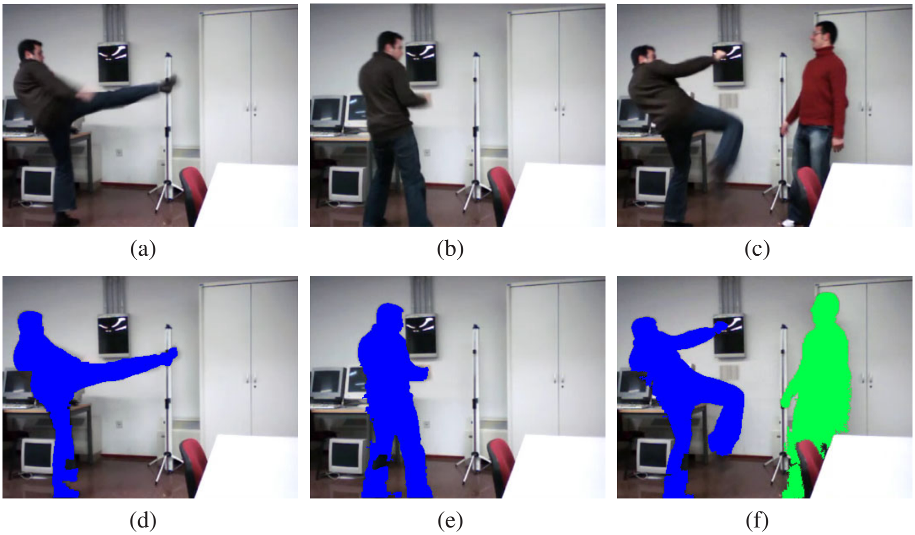

Fig. 4. Final results of the proposed method. In (a), (b) and (c) we can observe three frames of a scene in raw form. (d), (e) and (f) show the segmentation results using our neural approach (NCM+), after applying shadow detection and morphological operations.

A quantitative comparison among the different algorithms is shown in Table 1. 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 inform the number of misleading foreground pixels. The success rate, indicating the accuracy of the corresponding algorithm, is presented in the column labelled '% Matching'. It can be observed that our method gets better results than the rest of analyzed methods, although the false negative rate is not as good as in the MoG model. Nevertheless, our model defines more clearly the shape of the object. Therefore, making a post-processing phase using morphological operators, we can achieve a smaller error rate.

Final results of our method, showing a quite good segmentation, can be observed in Fig. 3. If these results are to be used in a subsequent tracking phase, others post-processing mechanisms are needed to achieve reliable solutions:

- Objects in motion in the scene usually can cast shadow on the background which could affect to the overall segmentation as occurs in Fig. 3. In this case, shadow pixels should be detected. Some techniques have been proposed in the literature [14,15,16,17] to this end. In our case, the method proposed in [18] has been implemented, based on the proportionality property between the shadow and the background in the RGB color space.
- As mentioned before, the results have also been improved by applying morphological operators after the segmentation process.

Figure 4 shows the final result of our approach, after applying a shadow detection method and enhancing the obtained frame by using morphological operations. As we can observe, this segmentation is effective enough to be used in a subsequent tracking phase.

Another advantage of the system is its efficiency (in terms of computational cost) compared with other techniques. It is possible to develop our neural network approach in a parallel hardware implementation, in which the segmentation time can be improved using a typical field programmable gate array (FPGA). Therefore, it will require a lower number of iterations to achieve the overall segmentation. Thus our method is able to perform in real-time.

4 Conclusions and Future Work

In this work, we have developed a new competitive neural model focusing on object detection and segmentation in video scenes. This new model is based on an unsupervised learning performed to model the RGB space of each pixel.

The main contribution of this model is the use of adaptive neighborhoods that incorporate some mechanisms of repulsion and consciousness. The neighborhoods used in the model are combination of crisp and fuzzy ones. The crisp neighborhood tries to model the vicinity of the synaptic weight vector in the RGB space, whereas the fuzzy one helps to avoid the apparition of dead neurons, i.e., neurons that never activate.

The segmentation accuracy of the proposed neural network is compared to mixture of Gaussian 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.

The proposed algorithm is 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, 2001, pp. 158–161 (2001)
4. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principles and practice of background maintenance. In: Proceedings of the 1999, 7th IEEE International Conference on Computer Vision (ICCV 1999), pp. 255–261 (1999)
5. Haritaoglu, I., Harwood, D., Davis, L.: W4: real-time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 809–830 (2000)
6. McKenna, S., Jabri, S., Duric, Z., Rosenfeld, A., Wechsler, H.: Tracking groups of people. Computer Vision and Image Understanding: CVIU 80(1), 42–56 (2000)
7. 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)
8. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1999)
9. Zabavin, N., Kuznetsova, M., Luk'yanitsa, A., Torshin, A., Fedchenko, V.: Recognition of handwritten characters by means of artificial neural networks. Journal of Computer and Systems Sciences International 38(5), 831–834 (1999)
10. Hall, L., Bensaid, A., Clarke, L., Velthuizen, R., Silbiger, M., Bezdek, J.: A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain. IEEE Transactions on Neural Networks 3(5), 672–682 (1992)
11. Pham, D., Xu, C., Prince, J.: Current methods in medical image segmentation. Annual Review of Biomedical Engineering 2, 315 ± (2000)
12. Mérida-Casermeiro, E., López-Rodríguez, D., Galán-Marín, G., Ortiz-de Lazcano-Lobato, J.M.: Improved production of competitive learning rules with an additional term for vector quantization. In: Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B. (eds.) ICANNGA 2007. LNCS, vol. 4431, pp. 461–469. Springer, Heidelberg (2007)
13. 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)
14. Cucchiara, R., Grana, C., Piccardi, M., Prati, A., Sirotti, S.: Improving shadow suppression in moving object detection with hsv color information. In: IEEE Intelligent Transportation Systems Conference Proceedings, pp. 334–339 (2001)
15. Mikic, I., Cosman, P., Kogut, G., Trivedi, M.: Moving shadow and object detection in traffic scenes. In: International Conference on Pattern Recognition, ICPR. IEEE Computer Society, Los Alamitos (2000)
16. Martel-Brisson, N., Zaccarin, A.: Learning and removing cast shadows through a multidistribution approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(7), 1133–1146 (2007)
17. Prati, A., Cucchiara, R., Mikic, I., Trivedi, M.: Analysis and detection of shadows in video streams: a comparative evaluation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, pp. 571–576 (2001)
18. Horprasert, T., Harwood, D., Davis, L.S.: A statistical approach for real-time robust background subtraction and shadow detection. In: Proceedings of IEEE International Conference on Computer Vision (1999)