Analysis of Perceptual Models Based on Visual Cortex for Object Segmentation in Video Sequences

Juan Alberto Ramirez-Quintana, Mario Ignacio Chacon-Murguia

Visual Perception Applications on Robotic Lab
Chihuahua Institute of Technology, Chihuahua, Chih, Mexico
jaramirez@itchihuahua.com, mchacon@ieee.org

Abstract. Visual perception capacities provide us with the ability to recognize objects from the interpretation of shape, color, orientation and motion features. The mechanisms in the visual cortex that allow the interactions between those visual features have been formalized in neurocomputational models and Artificial Neural Networks. In this paper, we propose a method based on perceptual models of visual cortex to analyze color, texture and motion in video sequences oriented to object segmentation. The results of the methods inspired in the behavior of the visual cortex have shown coherent object segmentation in videos with real scenes.

Keywords. Receptive fields, Visual cortex, video segmentation.

1 Introduction

Visual perception is a set of capacities of the brain that allows for interpretation of the information generated by the light reflected on object surfaces. These capacities lead to divide a scene in visual elements, with the aim to recognize coherent objects, using features like color, shape, orientation and motion [1]. There are many neurocomputational models and Artificial Neural Networks (ANN) based on the behavior of the visual cortex that attempt to explain the perception process. These models could be classified into three kinds: Adaptive Resonance Theory (ART), Pulse Neural Networks (PuNN) and Self-Organized Models (SOM). Among the ART-based models, there are some inspired in the behavior of different types of neurons in the visual cortex and their interaction between object recognition [2] and motion [3]. Pulse Networks can be classified according to their dynamic mechanisms of coupled oscillations [4] and integrate and fire networks like Spiking Neural Networks (SNNs) [5]. Coupled oscillations are models involving perceptual theories that describe the nervous system as a set of oscillations with a temporal dynamic that form groups of neurons. The Laterally Excitatory Globally Inhibitory Oscillator Network (LEGION) is a network used in image segmentation and scene analysis based in coupled oscillations [4][6]. The SNNs are dynamic models inspired in the membrane

© J. A. Olvera-López et al. (Eds.)
Special Issue: Advances in Pattern Recognition
Research in Computing Science 61, 2013, pp. 43-52
Paper Received 19-02-2013 and Accepted 22-04-2013

potentials of the neurons and they have been used to model different perceptual mechanism like in [7][8]. SOM models are inspired from the hypothesis that the synaptic weights in visual cortex are organized through self-organization based on the input stimuli. The most popular model of the SOM networks is the Kohonen Network, which have been used in several applications for pattern recognition and simulation of perceptual mechanism given in the visual cortex [9]-[11].

Visual perception theories have been a useful inspiration to develop methods for pattern recognition applied in computer vision, because it provides insights that allows the recognition of visual patterns in complex scenarios like the illusions shown in figure 1. Many works based on different approaches like ANNs or probabilistic methods try to relate the visual perception theories with image processing and computer vision applications. However, some of these works tend to be very specific in the application of certain principles, or they use unreal scenes for testing, and the processing times are not plausible for real time video analysis. Therefore, in this work we propose an analysis of neurocomputational models based on visual cortex theories and ANN, with the aim to design a real-time method using visual perception theories supported by visual cortex models for static or moving object segmentation in video sequences in real scenarios.

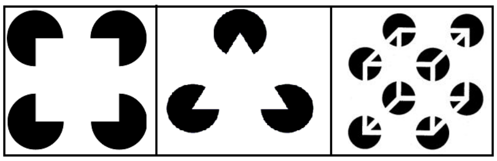

Fig. 1. Common illusions analyzed in visual perception theories.

We first present a state of the art of models based on the visual cortex that explain perceptual mechanism in object detection and motion. Then, we simulate the perceptual mechanism given in those models with ANN, and based on the simulations we describe a real-time method to perform coherent object segmentation in video sequences.

The rest of the paper is organized as follows: Section II presents the fundaments of the visual cortex. Section III analyses the models used in the segmentation process. Section IV describes the main aspects in the design of the segmentation methods and section V reports the final discussion.

2 Fundaments of Object Detection in the Visual Cortex

The visual cortex is the section of the brain that implements the main tasks of the visual perception process. This process starts in the retina, which acquire the color information with the cones, and the illumination changes with the rods. These light signals are converted in electrical signals, and they are distributed in the Retina Ganglia Cells (RGC) in a set of Receptive Fields (RF) calling ON and OFF. The RGCs have a set of axons that form the optic nerve and they are extended to the Lateral Geniculate Nucleus (LGN), which pass the information to the visual cortex and filter the useless information. The RFs are sensible to changes light-dark, wavelength Red-Green (R/G), Green-Red (G/R) and Blue-Yellow (B/Y) [12][13], all distributed in the channels ON, OFF, as illustrated in figure 2. They are modeled with Difference of Gaussians functions, DoG, where ON has a positive center Gaussian with variance less than the variance of the negative Gaussian, OFF has a negative center Gaussian with a smaller variance than the variance of positive Gaussian.

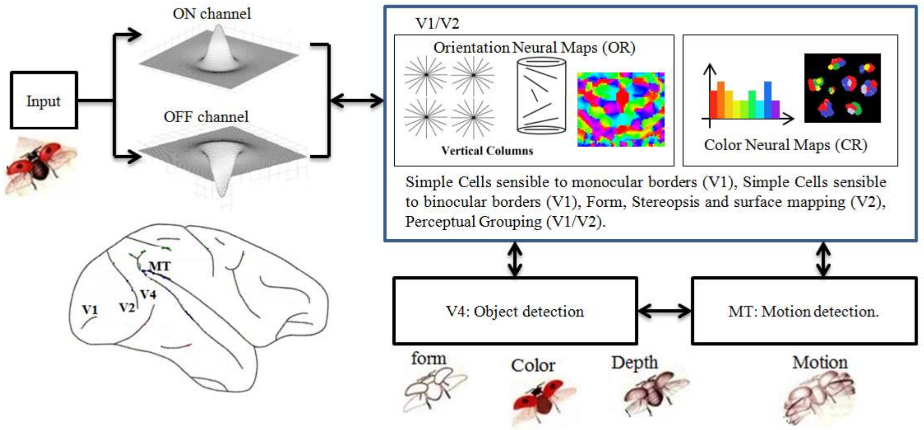
ON channel
Input
OFF channel
V1/V2
Orientation Neural Maps (OR)
Vertical Columns
Color Neural Maps (CR)
Simple Cells sensible to monocular borders (V1), Simple Cells sensible to binocular borders (V1), Form, Stereopsis and surface mapping (V2), Perceptual Grouping (V1/V2).
MT
V4
V1
V2
V4: Object detection
MT: Motion detection.
form
Color
Depth
Motion

Fig 2. Visual cortex parts for object and motion detection.

In the primary visual cortex (V1) and the visual area 2 (V2), the information of the RF is mapped in a set of neurons that are sensible to certain visual features. A common example of these neurons are the vertical columns [14][15], which are cells that describes the orientation patterns of the visual information under a retinotopic structure, and they are calling Orientation Map (OR). V1 and V2 have other similar maps sensible to color features (CR map) and ocular dominance (DO map) [16]. Those maps avoid the redundancy and develop the perception process forming features like shape, color, illumination changes, orientation and depth. According to the literature, it is consistent that features like color, orientation and depth are the inputs to the visual area 4 (V4), which is associated to the object detection and the medial temporal cortex (MT) is associated to motion perception. As figure 2 shows, the visual cortex has feedback loops in all its parts, which is an important item in the formation and recognition of visual patterns. The visual cortex has other cortical parts not analyzed in this work.

3 Scheme of the methods obtained from the perceptual models.

Visual perception of dynamic objects by the visual cortex involves a feedback between object detection and motion perception (V4 and MT). Therefore, with the analysis of those visual cortex parts, we propose the designing of a method based on the interactions between V1, V2, V4 and MT to achieve object segmentation, where the objects are classified in static and dynamics. The objects that can change its positions in the scene are dynamic objects, while those objects that remain in the same position in the scene during the video sequence are the static objects. The figure 3 shows a scheme of the proposed method, where the LGN processes the information in the RF channels with DoG. In the next step, V1/V2 processes different features used in the modules V4 and MT achieves the objects segmentation function. The videos used in the experiments were acquired with stationary cameras. The next subsections describe the visual cortex models used to design the segmentation methods.

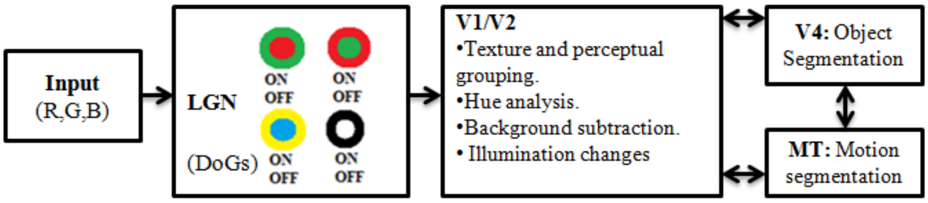

Fig 3. Scheme of the method to static and dynamic object segmentation.

3.1 V1 Feature extraction model

According to [12], V1 and V2 extract different features from visual patterns, using a set of neurons that organize the weights to form the cortical maps OR, CR, DO and illumination changes. To represent those cortical layers, we analyzed a set of models of the family of Laterally Interconnected Synergetically Self-Organizing Map (LISSOM) [15], which are developed from the theories on visual cortex behavior. The aim of LISSOM is to simulate the simple cells of the retina, LGN, V1 and in occasions V2 that activate the cortical maps. The basic LISSOM model (Fig. 4) is used to train the OR map, which describes the main orientations of a visual stimulus through time. The map is codified with hue levels like in the figure 4.

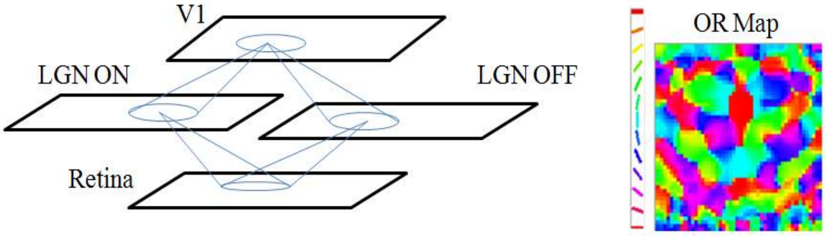

Fig 4. Architecture of the LISSOM model to activate the OR map, which is codified in the hue space according to the orientation bar shown.

Figure 5 shows a simulation of the LISSOM model, where the retina has two Gaussian patterns, the LGN ON/OFF have fixed weights given by DoG and the Figures 5b y 5c show the response of the LGN, which have afferent weights connected with V1. The cortical activation of V1 is based on the afferent weights, the lateral excitatory weights *E* and the lateral inhibitory weights *L*. The hebbian learning is used to update all the weights. The complete model is documented in [15]. The weights converge after approximately 10000 iterations (Figures 5e-5h). The weights model the OR map, V1 and both cortical maps are combined forming the response shown in figure 5d where V1 activates the orientation patterns of the visual stimuli in the actual iteration.

From among the LISSOM models, the Tricromatic LISSOM model (TLISSOM) is a model that describes how V1 and V2 stimulate the organization of the neural color map (CR). Another LISSOM model is Perceptual Grouping LISSOM (PGLISSOM), which is a set of SNNs that construct the OR map through pulse mechanism. This model drives the weight like the basic LISSOM model and solves perceptual grouping problems like the *kanizsa* triangle [15].

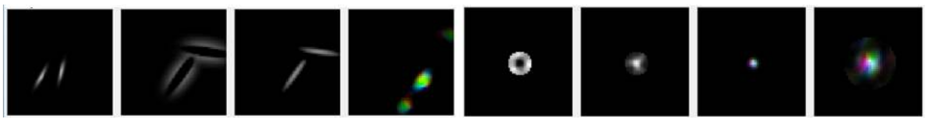

Fig. 5. LISSOM model response. (a). Input patterns, LGN response (b). OFF and (c). ON. (d). V1 final response. (e), (f). Afferent weights. (g). *E* weights. (h). *L* weights.

Among the LISSOMs models, the SOM Retinotopic Map (SOMRM) [15], represents the cortical mechanism of LISSOM, but SOMRM is similar to the *Kohonen* network. The figure 6a shows the architecture of the SOMRM, which have two layers; the retina and V1. The retina is a set of *K* receptors that sends the input pattern to V1, which is a set of neurons with synaptic weights in the interval of [0,1]. The neurons W_{ij} are represented by vector weights $W_{ij}=\{\omega_{1ij}, \omega_{2ij}, \omega_{kij} \ldots, \omega_{Kij}\}$ and according to [15], the initial response (η_{ij}) is given by the dot product between the neurons W_{ij} and the input I_k. The competition process to select the winner neuron in the SOMRM is given by the highest value in η_{ij}. Moreover, this competition process is different to the selection of the winner neuron in the Kohonen network, where the winner is given by the minimum value of the Euclidian distance between the input and the weights. According to [15], the learning of the SOMRM is based on the normalized hebbian rule determined as:

$$\omega_{k,ij}^{t_s+1} = \left(\omega_{k,ij}^{t_s} + \alpha I_k \beta_{ij}\right) \Big/ \sqrt{\sum_k \left(\omega_{k,ij}^{t_s} + \alpha I_k \beta_{ij}\right)^2} \tag{1}$$

where, t_s are the iterations of the SOMRM, α is the learning rate, which fall to zero if $t_s \to \infty$. β_{ij} is the neighborhood function computed as:

$$\beta_{ij} = \eta_{i_w j_w} \exp\left(-\frac{(i_w - i)^2 + (j_w - j)^2}{\sigma_\beta^{\ 2}}\right) \tag{2}$$

where (i_w,j_w) is the index of the winner neuron, σ_β is the neighborhood radio which falls to 0.5 if $t_s \to \infty$ [15]. Then, the SOMRM has two learning parameters; α, that defines the learning of all neurons in the hebbian rule and β_{ij}, that defines the learning capability per neuron based on the distance of each neuron to the winner.

The input pattern in the retina is a Gaussian that changes its position in each iteration like in figure 6b. In initial conditions, the weights in the SOMRM has random values in the interval [0,1] like in figure 6c, also $\alpha > 0$ and $\sigma_\beta > 0$. If $t_s \to \infty$, the neurons tend to learn the input pattern I_k, i.e., the neurons learn the Gaussian pattern on different positions as illustrated in figures 6d y 6e. Therefore, the SOMRM network is a simulation of V1 based on SOM, where each neuron develops capability of spatial selectivity, forming topographic maps in order to represent a pattern that changes its position in the retina surface.

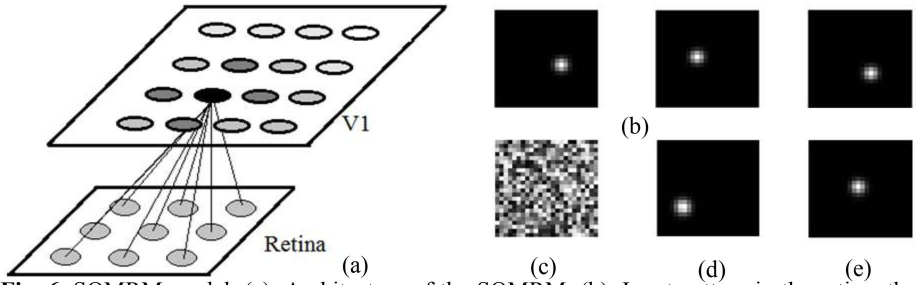

Fig. 6. SOMRM model. (a). Architecture of the SOMRM. (b). Input pattern in the retina, that changes its position. The size of the retina is *K=576* (*24x24*). (c). Neuron initial condition weights. (d) and (e). Neurons in the final iterations, where (d) is a neuron positioned in the border and (e) is a neuron positioned in the center.

3.2 Feature extraction for object detection

As we mentioned before, neurons in V1 and V2 are sensible to features of color, orientation and illumination changes. According to [13], the self-organizing interactions in V1 develop the modeling of the maps OR, CR.

The color analysis for the proposed scheme is based on TLISSOM model documented in [13]. This model describes how the training of the neural map CR is performed by the self-organization of V1, V2, and the processing of color RFs G/R, R/G and B/Y. Therefore, the method for color analysis in the proposed method is based on color RFs and cortical layers with self-organized process. The input of the method is a frame of a video sequence in the RGB color space. The RFs are represented with the sum of two convolutions between a Gaussian with a color channel. For example RF G/R ON is represented from the sum of convolutions between a positive center Gaussian with the red channel (R) and a negative surround Gaussian with higher variance with the green channel (G). In the channel G/R OFF, the sign of the center Gaussian is negative, and the sign of the surround Gaussian is positive. Despite the input of RFs is a RGB frame, the RF response is sensible to Hue, Saturation and Value (HSV) information of the frame. For example, as figure 7, the RFs are variant to H and S, but invariant to V changes (the response of the RF to the images in figures 7a y 7b are the same). The RF response is combined in V1, which is a self-organized cortical layer represented with an ANN based on SOMRM to produce four color channels (Red-Yellow-Green-Blue), all invariant to V changes. Those color channels are analyzed with their histograms to simulate the neural color map CR, which represent the activations in V2 that form the color groups. This method is inspired on TLISSOM that forms the color groups combining CR with the SOMRM and the color RFs is calling TRSOM and it achieves the object segmentation function using texture information.

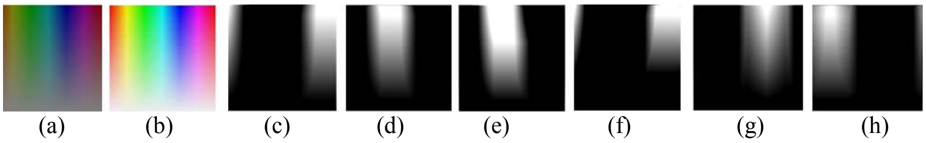

Fig. 7. RF Activation of the images with low bright (a), high bright (b). (c) and (d) RF Channels G/R. (e) and (f) Channels R/G.(g) and (h) Channels B/Y.

The texture and shape information is commonly analyzed with the orientation feature. Therefore, the proposed scheme uses the mechanisms focused to develop the orientation feature in the visual cortex for texture-shape analysis. For this feature, we used pulsed neural networks (PuNN), because they are based on the visual cortex behavior and they are commonly applied in image segmentation. From among the PuNN, the models LEGION [4][6] and PCNN [17] have been successfully applied in scene and texture segmentation. Therefore, we used the PCNN and the pulses were codified in time signatures defined in [17] which can perform the same neuronal synchronization as the coupled oscillations. Figure 8c illustrates an example of image segmentation with the PCNN defined in [17], where the time signature is calculated as:

$$SY_{ij} = \sum_{n=1}^{N} Y_{ij}(n) \tag{3}$$

where SY_{ij} is a time accumulation of the output pulses $Y_{ij}(n)$ in each iteration, N is the final iteration. Then, based on the PCNN time signatures and the PGLISSOM pulse mechanism [15], we design a pulsed model based on PCNN, where the input is a grayscale image. The RFs ON and OFF are given by DoG and the response of LGN is the convolution between the input and each RF channels. Then, LGN response pass to V1, which is represented by a set of Orientation Filters that represent the vertical columns in OR. To simulate the membrane actions potential in V1 we use the time signatures of the equation 3, where the pulses were modeled with a PCNN[17]. Figure 8d shows the proposed model, denominated RF-PCNN. In order to test the plausibility of the model with visual perception theories, we used two images commonly analyzed in the perception literature. Figure 8e shows an image with the *Hermann* illusion, where there is an optical illusion characterized by false blobs in the intersections of the grids. The strong suppression of the ON channel in the intersections causes this illusion [18]. Figure 8f shows an illusion where the observer could find a word if he/she tries to shrink the eyelids. When the RF-PCNN processes both images, the model obtains output patterns consistent with the perceptual illusions, as shown in figures 8g and 8h. Thus, there is evidence to assume that the model is consistent with the perceptual mechanisms. Following this conclusion and knowing that the PCNN has been used for object segmentation and that the RF-PCNN develops perceptual mechanism similar to the neurocomputational model PGLISSOM[15], the RF-PCNN model is proposed as the basis for the method of texture and shape segmentation. Thereby, V1/V2 extracts the features of texture shape with the PCNN and color with the TRSOM. Both features should be combined in a method inspired by the operation of V4 for complete the static object segmentation.

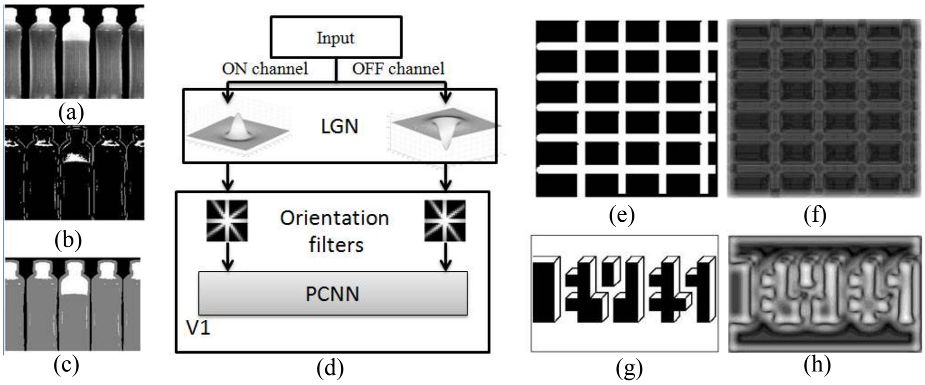

Fig. 8 PCNN model. (a). Input image for segmentation using PCNN time signatures. (b). Pulse in n=15. (c). Segmentation of (a) given by eq. (3). (d). RF-PCNN model. (e). Input based on the *Hermann*. (f). Output of the model RF-PCNN with *Hermann* illusion. (g). Image with hidden word '*eyes*'. (h). Output of the model RF-PCNN with hidden word 'eyes'.

4 Design of object segmentation methods

In the case of static object segmentation, we are developing a method inspired in V4, which is based on features of texture and color. The proposed scheme obtains the color feature from the segmentation generated by the TRSOM, and the texture features from the segmentation inspired by the RF-PCNN model. The inputs to this models are the V information defined as the maximum value in a RGB pixel and the neighborhood information of each pixel. The figures 9b and 9e shows the results of the both segmentation process.

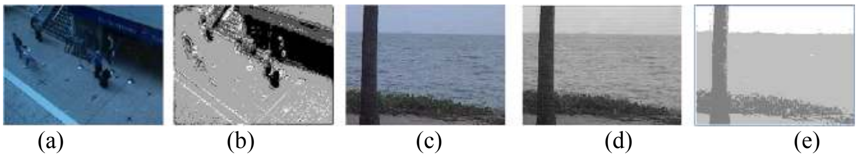

Fig. 9. Color and texture analysis in video sequences. (a). Frame of the video'*14_p2007_3*'. (b). Color segmentation. (c). Frame of the video '*WS*'. (d). V information of (c). (e). Texture segmentation of (d).

With respect the dynamic object segmentation, we design a method based on background subtraction approach to extract all the motion regions, which is defined as visual alert. Also, with the aim to find the dynamic objects, we included an analysis of spatial and temporal illumination changes, which are defined by an accumulation of the differences between consecutive frames in temporal sequences. This accumulation forms a spatiotemporal pattern that describes the motion of dynamic objects. This spatiotemporal pattern represents the process in V1 to find motion patterns [3].

The background subtraction (BS) method is based on topographic interactions in V1, therefore we designed a model inspired in SOMRM for background modeling. The motion detection is based on BS and a model involving Cellular Neural Networks to reduce noise. Both, background subtraction and the illumination changes are correlated to find the dynamic objects in a method inspired in the behavior of MT. Figure 10 shows a scheme of the proposed method for dynamic object segmentation along with a comparison of a segmentation and its groundtruth.

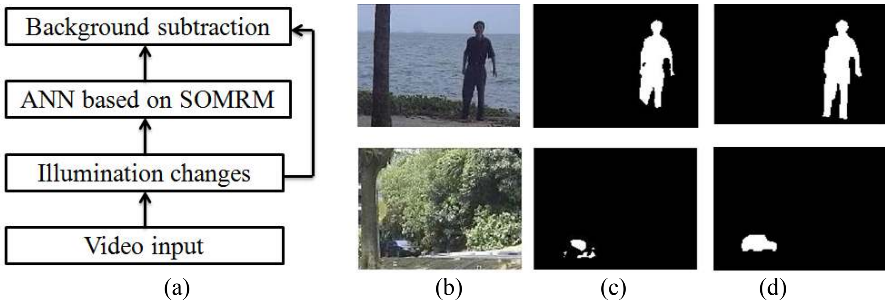

Fig. 10. Dynamic object segmentation method performs. (a). Scheme of the method. (b). Two video frames. (c). Dynamic object segmentation. (d) Groundtruth.

The validation of the dynamic object segmentation method was based on the metrics of Precision (P), Recall (R), F1and Similarity (S), all defined in [19], because these metrics are widely used in the literature related to dynamic object segmentation. The videos used were obtained from databases used in the literature and the methods were tested in critical condition. Table 1 presents the results which indicate a good segmentation performance except when dynamic objects have chromatic features similar to the background. The videos have a resolution of 120x160 and the dynamic object segmentation method run at 40 frames per second (fps). The videos were obtained from databases wallflowers and perception, which have been used to validate different dynamic segmentation methods.

Table 1. Dynamic Object Segmentation Results.

Video	CF	MR	FT	LB	MO	TD	WS	WT
P	0.9687	0.9254	0.875	0.9556	1	0.7876	0.9871	0.9769
R	0.9291	0.9177	0.7794	0.9188	1	0.7546	0.8328	0.9074
F1	0.9485	0.9215	0.8244	0.937	1	0.7707	0.9034	0.941
S	0.902	0.9215	0.7013	0.8811	1	0.627	0.824	0.889

5 Final Discussion

In this paper we have presented the progress in the implementation of a bioinspired method for static and dynamic object segmentation in video sequences, which is based on perceptual models of the visual cortex. This method was proposed by assuming that object detection is given by the analysis of features like color orientation and depth, while motion detection is based on spatiotemporal analysis of illumination changes and visual alert. From these models, we designed different methods to extract features of color, texture, alert and illumination changes, which were used to develop a set of methods that form a neuroinspired object segmentation scheme. Findings showed that the object segmentation is coherent except when the color features of the objects are very similar between them. For this reason, future work we will be oriented to finish a method for static object segmentation inspired in V4 and we will develop a feedback scheme as in the visual cortex models to improve the perception process in the object segmentation.

6 References

1. Schwartz S.: Visual Perception a clinical orientation. McGraw Hill (2010) 1-3.
2. Cao Y. and Grossberg S.: Laminar Cortical Model of Stereopsis and 3D Surface Perception: Closure and da Vinci Stereopsis. Technical Report, (2004).
3. Berzhanskaya J., Grossberg S., Mingolla E.: Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception. Technical Report, (2007).
4. Yu G. and Slotine J.: Visual Grouping by Neural Oscillator Networks, IEEE trans on Neural networks Vol. 20, No 12, pp 1871 – 1884, (2009).
5. Izhikevich E.: Which Model to Use for Cortical Spiking Neurons?, IEEE trans on Neural networks, vol 15, No 5, pp 1063-1070, (2004).
6. Benicasa A., Quiles M., Zhao L. and Romero R.: An Object-Based Visual Selection Model with Bottom-up and Top-down modulations; Brazilian Symposium on Neural Networks, pp 238-243, (2012).
7. Koene R., Hasselmo M.: An integrate and fire model of prefrontal cortex provides a biological implementation of action selection in reinforcement learning theory that reuses known representations, Int. Conf. on Neural networks, pp 2873-2878, (2005).
8. Ratnasi S. and McGinnity T.M.: A Spiking Neural Network for Tactile Form Based Object Recognition; Int. Join Conf on Neural Networks, pp 880-885, (2011).
9. Baier V.: Motion Perception with Recurrent Self-Organizing Maps Based Models, Int. Joint Conf. on Neural Networks, 1182-1186, (2005).
10. Kiang M.: Extending the Kohonen self-organizing map networks for clustering analysis, Elsevier Computational statistic & data analysis, Vol 38, No 2, pp 161-180, (2001).
11. Chacón-Murguia M. I. and Gonzalez-Duarte S.: An Adaptive Neural-Fuzzy Approach for Object Detection in Dynamic Backgrounds for Surveillance Systems, IEEE Trans. on Industrial electronics, Vol. 59, No. 8, , pp. 3286-3298, (2012).
12. Kruger N., Janssen P., Kalkan S., Lappe M., Leonardis A., Piater J., Rodriguez-Sanchez A., Wiskott L.: Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol PP, No 99, 1-24, (2012).
13. Ben De Paula J.: Modeling the self-organization of color-selective neurons in the visual cortex, Report AI-TR-07-347, (2007).
14. Yu B., Zhang L.: Pulse-Coupled Neural Networks for Contour and Motion Matchings, IEEE trans on Neural networks, Vol. 15, No 5, pp 1186-101, (2004).
15. Miikkulainen R., Bednar J., Choe Y., Sirosh J.: Computational Maps in the Visual Cortex, Springer Sciences Media Inc, (2005).
16. Bednar J.: Building a mechanistic model of the development and function of the primary visual cortex, Journal of Physiology, pp 194-211, (2012).
17. Wang Z., Ma Y., Cheng F., Yang L.: Review of pulse-coupled neural networks, Image and Vision Computing, Vol 28, No 1, pp 5–13, (2010).
18. Frisby J. and Stone J., Seeing, The computational approach to biological vision, MIT Press, (2010).
19. Fan-Chieh C., Shih-Chia H. and Shanq-Jang R., “Illumination-Sensitive Background Modeling Approach for Accurate Moving Object Detection”, IEEE Trans. on broadcasting, Vol 57, No 4, pp 794-801, (2011).