21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Enhancing Object Detection Performance by Integrating Motion Objectness and Perceptual Organization

Concetto Spampinato, Simone Palazzo
Department of Electrical, Electronic and Computer Engineering - University of Catania
Viale Andrea Doria 6 - 95125, Catania -Italy
{*cspampin, palazzosim@dieei.unict.it*}

Abstract

In this paper we propose a method to improve the performance of motion detection algorithms by estimating the probability that a detected blob (i.e. a group of pixels identified as foreground) is actually an object of interest. The system exploits "objectness" and perceptual organization to estimate general properties of real-world objects such as convexity, symmetry, well-defined boundary, visual contrast and cohesiveness. The measures of these properties are given as input to a naive Bayes classifier, which is trained to distinguish objects of interest from false positives. The system was trained and tested on two "real-life" environments (underwater and vehicular monitoring) and the results showed an increase of the performance of four state-of-art motion detection algorithms of about 15%. We also tested our approach on the CAVIAR dataset and although the system was not trained on that specific object class (people) it was able to increase the object detection performance of about 10%.

1. Introduction

In the last decades, object detection in video streams has become one of the most active research area in image processing, and in general in computer vision. The problem has been mainly tackled by background subtraction techniques, which compare the observed image with an estimated model (referred to as the background) describing the scene without objects of interest. A myriad of motion detection algorithms has been proposed for discriminating background from foreground [1, 2] although none of them have demonstrated to be generally superior to the other ones, and the the false positive problem is still unsolved. This leads to the need of a side processing level exploiting generic object features to repair failures that may naturally happen during the background/foreground classification process. In the literature, there exist approaches, such as the traditional salient blob detectors [3] that use features of arbitrary object classes to estimate the probability that a certain window in a test image contains an object of interest. Recently, Bogdan *et al.* in [4] used the "objectness", assessed as the probability of a window within an image to contain an object measured by integrating several image cues, to improve the performance of class specific object detectors on a set of very complex objects of the PASCAL VOC 07[1]. However, these approaches are thought to detect objects in static images by evaluating iteratively a huge number of windows and as such they cannot be applied to process video stream for two main reasons: 1) they limit their analysis only to what is observable in a single frame, ignoring motion information 2) they do not adopt general properties of real-world objects. In order to overcome the existing approaches and to enhance motion detection performance in this paper we propose a method that assesses the probability that a moving blob be an object of interest by integrating general and specific features of real-world objects. The contributions of our paper are: 1) a human perceptual organization model to discriminate blobs that are most likely produced by the motion of a biological object from blobs that may arise due to changes in the background (i.e. luminosity), 2) we extend the concept of "objectness" to "motion objectness" to describe the probability that a blob moving in a video stream is a specific class object and 3) we improve significantly the performance of state-of-art detectors.
The remainder of the paper is as follows: Sections 2 and 3 describe the adopted perceptual organization model and the "motion objectness", whereas experimental results and concluding remarks are discussed, respectively, in Section 4 and 5.

[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

2. Perceptual Organization Model

The ability of humans to identify objects, and more in general structures, without a priori knowledge on their contents is referred as perceptual organization, which, according to Gestalt laws [5], is led by some basic principles such as proximity, similarity, continuity, symmetry and convexity. To measure quantitatively the Gestalt laws in real-word applications, we have adopted the method proposed in [6] which encodes such laws into a boundary energy function:

$$E\left[\partial R\right] = \frac{-\int\int_R f\left(x,y\right) dxdy}{L\left(\partial R\right)} \tag{1}$$

where ∂R is the object's contour, $L\left(\partial R\right)$ the contour's length, and $f\left(x,y\right)$ is a weight function for each point belonging to the object. The first step for the evaluation of $f\left(x,y\right)$ is a superpixel segmentation [7] of the object's region into homogenous patches. For each pixel (x,y), belonging to patch i, the corresponding weight is computed as:

$$f\left(x,y\right) = e^{-\theta \bullet \eta(S_i - S_a)} \tag{2}$$

In this formula, S_i is a two-component vector $[B_i \;\; C_i]$, where B_i and C_i represent respectively the boundary complexity of patch i and its cohesiveness with the other patches which make up the object. S_a is a reference vector computed on the largest patch of the object [6]. θ is a weight vector and η is vector element-by-element absolute value.

3. Motion Objectness

"Motion Objectness" is defined as the probability that a detected blob be an object of interest. The intuition behind this concept is that any object of interest is featured by specific intraframe and interframe properties with respect to a background object.

3.1 Intraframe Properties

To describe the peculiarity of an object within a image, we have adopted the following characteristics:

- *Closed boundary in space*. This property aims at evaluating how the blob's contour matches the object's boundary. To measure it, we assess the density of edges included in a blob. Let δ_b and b^θ_{in} be, respectively, the contour of the considered blob b and the inner blob obtained by shrinking b of a factor θ. The edge density is given by:

$$ED = \frac{\sum_{p \in b^\theta_{in}} M_{ED}(p)}{A_{b \backslash \delta_b}} \tag{3}$$

where $M_{ED}(p)$ is a binary edgemap and indicates if the pixel p is classified as edge by a Canny edge detector. $A_{b\backslash\delta_b}$ is the area of the blob b minus its contour δ_b.
Moreover, we have also used the percentage of superpixels intersecting the blob contour computed as follows:

$$SI = 1 - \frac{\sum_{s \in \mathcal{S}} min(|s \backslash b|, |s \bigcap b|)}{|b|} \tag{4}$$

where $\mathcal{S}$ is the set of superpixels achieved with the algorithm proposed in [7] and $|b|$ the blob's area.

- *Appearance different from surrounding areas*. The dissimilarity of an object to its surrounding area is estimated by analysing color contrast along the object's boundary. Let b^θ_{out} and b^θ_{in} be the outer and the inner blob obtained, respectively, by dilating and shrinking the original blob b of a factor θ (empirically set to 2 in our implementation), the color contrast along the boundary of a blob b is computed as the Chi-square distance between the LAB histograms of the two rings (outer and inner) surrounding the object's boundary:

$$CC = \chi^2(h(b \backslash b^\theta_{in}), h(b^\theta_{out} \backslash b)) \tag{5}$$

- *Internal homogeneity of color and texture*. Most objects appear to have a limited number of colors and an uniform texture, especially when compared with complex background objects (e.g. trees, buildings, etc.). The internal homogeneity of a blob has been assessed by computing the average color value and the average texture of all the superpixels in a blob. The more similar these average results are, the more likely the detected blob is actually an object of interest. The average color homogeneity is given by the following formula:

$$HC = 1 - \frac{\sum_{s \in \mathcal{S}} \left|\left|C_s - \overline{C}\right|\right|}{Dim(\mathcal{S})} \tag{6}$$

where $\mathcal{S}$ is the same set of superpixels described above, C_s is the average color within each superpixel s, $\overline{C}$ is the average color within the whole blob and $Dim(\mathcal{S})$ the number of superpixels. Analogously, the measurement of the internal texture homogeneity is performed by averaging the outputs of a bank of Gabor filters applied to each superpixel in the detected blob.

- *Preferred positions*. The spatial coordinates of the blob's centroid are also used to measure "objectness", since we assume that some positions are more likely than others to contain objects.

Figure 1. Examples of the estimated probability of the windows to contain objects of interest.

3.2 Interframe Properties

Any object holds the *motion coherence* property that allows us to distinguish it from the rest of the scene. To measure this property, we have proposed two cues based on the object's motion vector (computed according to [8]):

- *Difference of motion vectors at object boundary*: Let $M_{b,in}$ and $M_{b,out}$ be the average motion vectors computed, respectively, in the ring just inside and the ring just outside of the blob's boundary, the motion difference at object boundary Δ_{MV} is assessed as the Chi-square distance between the motion histograms assessed in the two rings (size empirically set to 3 in our implementation) surrounding the object's boundary:

$$\Delta_{MV} = \chi^2(h(M_{b,out}), h(M_{b,in})) \qquad (7)$$

- *Internal motion homogeneity*. This cue is based on the assumption that the internal motion vectors of a correctly-detected object are more uniform than the ones of a false positive. The object is divided into a set of subpixels (as above) and the average motion vectors of each subpixel are compared. The internal motion homogeneity is computed as follows:

$$MH = 1 - \frac{1}{Dim(\mathcal{S})} \sum_{s \in \mathcal{S}} \frac{1}{Dim(\mathcal{V}_s)} | \sum_{p \in \mathcal{R}_s \cap \mathcal{V}_s} |M_p| - \overline{M}|^2 \qquad (8)$$

where $\mathcal{S}$ is the set of superpixels in the analysed blob and $Dim(\mathcal{S})$ its dimension, $\mathcal{R}_s$ is the union between the superpixel's current bounding box $R_s(t)$ and the bounding box of its last appearance $R_s(t-1)$, $\mathcal{V}_s$ is the set of valid points in $\mathcal{R}_s$, i.e. the points whose displacement project them inside $\mathcal{R}_s$ and $Dim(\mathcal{V}_s)$ is the number of valid points in $\mathcal{V}_s$. Finally, M_p is the motion vector (two components: x and y) describing the displacement of pixel p in two consecutive appearance and $\overline{M}$ is the average motion vector between all superpixels.

4. Experimental Results

The feature vector, containing the above objectness measures and POM energy value, of each detected blob is given as input to a naive Bayes classifier with two classes: *"object of interest" (OI)* and *"false positive" (FP)*, which computes the probability of the considered blob being an object of interest. We used 15 videos divided into three categories (of increasing difficulty) to train and test the Bayesian classifier: 1) 5 underwater videos targeting fish from the Fish4Knowledge[2] project's dataset, 2) 5 videos from the CAVIAR dataset[3], showing people in walking and meeting in closed environments (e.g. shopping centre) and 3) 5 videos of vehicles in a urban environment. The ground truths for CAVIAR videos were downloaded from their website, whereas the ground truths for the underwater and the vehicle videos were manually hand-labeled by us. The Bayesian classifier was trained using 6 videos (3 vehicle videos and 3 underwater videos). The samples corresponding to the OI class were obtained by computing the features described in Sections 2 and 3 on the ground truth objects; the samples used to train the FP class were computed on random blobs (not overlapping with correct object windows) taken from each video. In order to evaluate our approach's capability to distinguish objects of interest from false positives, we executed a set of state-of-the-art object detection algorithms (see Table 3.2) on the the remaining 9 videos. We did not include CAVIAR videos in the training set, in order to show that the proposed measures are generic over class. For each algorithm, we compared the results obtained by the algorithm itself with those obtained by excluding from the detected objects those with a OI probability lower than 50%. Table 1 compares the performance of each algorithm with and without the proposed method, in terms of detection rate DR and false alarm rate FAR. It is possible to notice how our evaluation method improves the performance of the

[2] http://fish4knowledge.eu
[3] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

	Vehicles				People				Fish			
	P		IP		P		IP		P		IP	
	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR	DR	FAR
AGMM [1]	83.5%	12.7%	83.2%	10.4%	72.1%	16.7%	69.1%	12.7%	66.9%	27.4%	64.8%	19.2%
APMM [2]	82.2%	10.6%	81.6%	7.7%	75.5%	18.8%	73.4%	13.9%	69.1%	26.7%	65.3%	20.6%
IM [9]	76.8%	13.9%	75.7%	9.4%	68.0%	22.3%	65.7%	16.0%	61.5%	32.3%	58.3%	16.3%
ViBe [10]	88.5%	10.1%	88.1%	8.0%	77.1%	15.6%	75.2%	12.2%	74.4%	19.1%	70.0%	12.9%

Table 1. Comparison between the performance of detection algorithms with and without our probability-based object filter. P: original performance of the algorithm, IP: increased performance

algorithms by drastically reducing the FAR, while DR remains unchanged. The results also show that the improvement introduced by our method is more evident for the underwater videos and less visible in the vehicle ones. This is due to the fact that false positives are mainly caused by moving background and object-background occlusions, which are much more frequent in underwater videos than in constrained environments. Fig. 1 shows some examples of probabilities computed on objects, background areas and occlusions.

5. Concluding Remarks

In this paper we propose a method for the estimation of the probability that an object in a video scene is a real object of interest, by exploiting real-word object properties such as convexity, visual contrast, well-defined boundary, symmetry and cohesiveness. This approach was trained and tested on underwater and urban videos using four state-of-art detection algorithms in order to analyze the improvement in the identification of false positives. Results showed an increase of the performance of the algorithms by 10%-15%, both for object classes used to train the classifier (vehicle and underwater environments), and for unseen objects (people). Moreover, these improvements come at an acceptable increase of the computation time for real-time 25-fps video processing.

6 Acknowledgements

This research was funded by European Commission FP7 grant 257024, in the Fish4Knowledge project (www.fish4knowledge.eu).

References

[1] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *CVPR'99*, vol. 2, no. c, pp. 246–252, 1999.

[2] A. Faro, D. Giordano, and C. Spampinato, "Adaptive background modeling integrated with luminosity sensors and occlusion processing for reliable vehicle detection," *IEEE Trans. on ITS*, vol. 12, no. 4, pp. 1398–1412, 2011.

[3] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *CVPR07*, pp. 1–8, 2007.

[4] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on PAMI*, vol. 99, no. PrePrints, 2012.

[5] Z. Liu, D. W. Jacobs, and R. Basri, "The role of convexity in perceptual completion: beyond good continuation.," *Vision Res*, vol. 39, no. 25, pp. 4244–4257, 1999.

[6] C. Cheng, A. Koschan, C.-H. Chen, D. L. Page, and M. A. Abidi, "Outdoor scene image segmentation based on background recognition and perceptual organization," *IEEE Trans. on Image Processing*, vol. 21, no. 3, pp. 1007–1019, 2012.

[7] P. Felzenszwalb and D. Huttenlocher *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[8] J.-Y. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker description of the algorithm," 2000.

[9] F. Porikli, "Multiplicative background-foreground estimation under uncontrolled illumination using intrinsic images," in *MOTIONS '05*, vol. 2, pp. 20 –27, jan. 2005.

[10] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Trans. on Image Processing*, vol. 6, no. 20, pp. 1709–1724, 2011.