Appl Intell (2014) 40:358–375
DOI 10.1007/s10489-013-0461-5

A machine learning based intelligent vision system for autonomous object detection and recognition

Dominik Maximilián Ramík · Christophe Sabourin · Ramon Moreno · Kurosh Madani

Published online: 29 August 2013

Abstract Existing object recognition techniques often rely on human labeled data conducting to severe limitations to design a fully autonomous machine vision system. In this work, we present an intelligent machine vision system able to learn autonomously individual objects present in real environment. This system relies on salient object detection. In its design, we were inspired by early processing stages of human visual system. In this context we suggest a novel fast algorithm for visually salient object detection, robust to real-world illumination conditions. Then we use it to extract salient objects which can be efficiently used for training the machine learning-based object detection and recognition unit of the proposed system. We provide results of our salient object detection algorithm on MSRA Salient Object Database benchmark comparing its quality with other state-of-the-art approaches. The proposed system has been implemented on a humanoid robot, increasing its autonomy in learning and interaction with humans. We report and discuss the obtained results, validating the proposed concepts.

D.M. Ramík · C. Sabourin (✉) · K. Madani
LISSI EA 3956, Senart-FB Institute of Technology, University Paris Est-Creteil (UPEC), Avenue Pierre Point, 77 567 Lieusaint Cedex, France
e-mail: christophe.sabourin@u-pec.fr

D.M. Ramík
e-mail: dominik.ramik@u-pec.fr

K. Madani
e-mail: kurosh.madani@u-pec.fr

R. Moreno
Facultad de Informática, Grupo de Inteligencia Computacional de la Universidad del Pais Vasco, Paseo Manuel de Lardizabal, 1 20018 San Sebastian, Guipuzcoa, Spain
e-mail: ramon.moreno@ehu.es

Keywords Intelligent machine vision · Visual saliency · Unsupervised learning · Object recognition

1 Introduction

The design of perceptual functions is a major problem in robotics. Fully autonomous robots need perception to navigate in space and recognize objects and environment in which they evolve. But the question of how humans learn, represent, and recognize objects under a wide variety of viewing conditions presents a great challenge to both neurophysiology and cognitive research, and of course, in the robotics field [9]. This paper focuses on a fundamental skill in artificial intelligence applied to robots evolving in a human environment, which are the learning and recognition process based on visual perception.

In the past decade, the scientific community has witnessed great advance on the field of techniques for object detection and recognition, such as SIFT [22], SURF [7], Viola-Jones framework [34], to mention only a few. More recently, in [16] Hossain et al. proposed "knowledge-based flexible-edge-matching" method to detect moving object detection in video sequences. Some works, also, have focused on feature selection and feature extraction like in [19]. Other works focused on cognitive approaches, as in [35], where authors have proposed to use a memory-based cognitive modeling for robust object extraction and tracking. Many of them were so successful, that we are already meeting them in commercial applications like cameras focusing automatically on human faces or product logo recognition in mobile applications. While these methods show often high rates of recognition and are able to operate in real time, they all rely on a human made databases of manually segmented or labeled images containing the object of interest without extensive spurious information and background. There are many

advanced works applying the mentioned approaches to enable learning and recognition in mobile robots, e.g. the Curious George robot described in [23], or [3] who use SIFT in context of a robust neural network based object recognition framework. Some of the techniques use such a database as learning samples to train a set of classifiers [34], others use it as a bank of templates for matching processes [6, 7]. The mentioned database is sine qua non for a successful recognition process, but its manual creation often requires a considerable time and a skilled human expert. This impedes design of a fully autonomous machine vision system, which would learn to recognize new objects on its own.

Motivated by the mentioned shortcomings regarding existing as well object recognition methods as vision systems, we present in this work an intelligent machine learning based system able to observe and learn autonomously individual objects present in a real environment. The goal for this system is to allow an embodied agent (e.g. a camera equipped mobile robot) to learn and to recognize objects encountered in its environment in completely autonomous manner. This addresses a foremost problem relating to autonomous robots' area, as in a more generic way autonomous artificial vision paradigm. It appears also as vital for reaching higher-level intelligent machines: machines able to "discover" and to learn unknown (e.g. not formerly memorized or stumbled upon) object from an unlabeled image. A clear need appearing as a first expected skill is the ability to select from the overwhelming sensory information (e.g. visual information) only the pertinent ones. Then, additional manifest needs are visual knowledge acquisition (learning the pertinent visual information) and already encountered objects' recognition (namely, based of acquired knowledge).

In design of such a system, our approach has been inspired partly by existing clinical investigations describing human vision system and partly by the way human learns objects. In fact, the extraction of objects of interest from raw images is driven by visual saliency. Building on existing work relating the field of visual saliency, we propose an intelligent vision system (concept, architecture and implementation on real robot) taking advantage from a novel salient objects' detection algorithm. Our choice to consider a spherical interpretation of RGB color space allows the system to take advantage from photometric invariants. This conducts to a fast image segmentation algorithm, robust to real-world illumination conditions, which serves to extract objects for learning from raw images. Resulting extracted objects can be then learned by most of the up to date machine-learning approaches in order to ease their recognition if they would be encountered afterwards. Concerning the implementation, two well known machine-learning based fast recognition methods have been tested: Speeded-up Robust Features (SURF) introduced in [6] and the Viola-Jones object detection framework, presented in [34]. Finally we validate our artificial cognitive system with an embedded color camera on a mobile robot in an explore-and-learn task performed in a common office environment.

The reminder of the work is organized as follows. Section 2, after a short overview of existing techniques about visual saliency, presents constitutive parts of our system and explains their interaction. In Sect. 3, we present our approach to saliency detection. In Sect. 4, we detail our salient object extraction technique. Learning procedures are detailed in Sect. 5. The Sect. 6 reports and discusses results of validation of the proposed intelligent vision system. Conclusion and further work perspectives are presented in Sect. 7.

2 Autonomous object detection and recognition based on visual saliency

In recent years, there has been a substantial progress in robotic systems able to robustly recognize objects in real world using a large database of pre-collected knowledge (see for example [3, 23]). But a fully autonomous robot cannot rely solely on a priori knowledge that has been given to it by a human expert. On the contrary, it should be able to learn on-line, in the place where it is used. If we want to allow a machine vision system to learn how to recognize an unknown object from an unlabeled image, we are in a similar situation as human parents teaching an infant. A clear need, for human and for machine as well, is from the overwhelming flow of sensory information to choose only the one which is pertinent in context of the given task. This is why, in our artificial visual cognitive system, the extraction of objects of interest from raw images is driven by visual saliency. Then, the resulting extracted objects can be exploited by existing recognition methods like SURF or Viola-Jones in our case. It is the ability of calculation of visual saliency, that enables our system to extract and learn individual objects and thus it is the cornerstone of our system. For this reason we will dedicate the following Sect. 2.1 to remind existing techniques about visual saliency. After this remember, Sect. 2.2 describes an overview of our cognitive system for visual perception.

2.1 Existing techniques about visual saliency

It may be generalized, that it is the saliency (in terms of motion, sound, color, etc.) that makes the pertinent information to be remarked or to "stand-out" from the context. For our purposes, we identify the mentioned explicitness with visual saliency. We argue that presenting an object in an explicit way, i.e. that it becomes visually distinct with respect to the rest of the scene may enable unsupervised extraction of such an object from the perceived image so that it can be used as a learning example for an object detector. We show further

on (see Sect. 6 for our experimental setup and images) that this explicitness does not necessitate an artificial scene setup (e.g. putting the learned object on a white background) but, on the contrary, that our system is able to learn objects presented in natural indoor conditions without any particular scene arrangement. This point is very important in the design of a fully autonomous vision system able to learn and recognize salient objects in a real environment.

Visual saliency (also referred in literature as visual attention, unpredictability or surprise) is described as a perceptual quality that makes a region of image stand out relative to its surroundings and to capture attention of the observer [2]. The inspiration for the concept of visual saliency comes from the functioning of early processing stages of the human vision system and is roughly based on previous clinical research. In early stages of the visual stimulus processing, human vision system first focuses in an unconscious, bottom-up manner, on visually attractive regions of the perceived image. The visual attractiveness may encompass features like intensity, contrast and motion. Although there exist solely biologically based approaches to visual saliency computation, most of the existing works do not claim to be biologically plausible. Instead, they use purely computational techniques to achieve their goal. One of the first works to use visual saliency in image processing has been published by [18]. Authors use an approach based on a center-surround contrast calculation using Difference of Gaussians. Other common techniques of visual saliency calculation published more recently include graph-based random walk [15], center-surround feature distances [1], multi-scale contrast, center-surround histogram and color spatial distribution [21] or features of color and luminance [2]. A less common approach is described in [20]. It uses content-sensitive hypergraph representation and partitioning instead of using more traditional fixed features and parameters for all images.

In image processing, identification of visually salient regions of an image are used in numerous areas including smart image resizing [5], adaptive image display on small device screens [10], amelioration of object detection and recognition [28], content based image retrieval and adaptive image compression or image collection browsing to mention only a few. Depending on the particular technique, many approaches like [1, 2] or [21] output the saliency map, which is an image whose pixel intensities correlate with the saliency of the corresponding pixels of the original image. Selection of the most salient regions from saliency map by application of a threshold or a segmentation algorithm is subsequently performed. It results into extraction of a visually important object or a patch of objects rather than just of a semantically incoherent fragment of the image. This property is exploited by several authors. In [8] a biologically-motivated saliency detector is used along with an unsupervised grouping algorithm to group together images containing visually similar objects. Notably in the work [31] a purely bottom-up system based on visual attention is presented, investigating the feasibility of unsupervised learning of objects from unlabeled images. Experiments are successfully conducted by its authors on real world high-resolution still images and on a camera-equipped mobile robot, where the capacity to learn landmark objects during its navigation in an indoor environment is shown. The main difference between this approach and the our one is, that Rutishauser et al. use visual saliency rather to indicate interesting parts of the input image, while we use it explicitly for extraction of individual visually important objects. Recently [12] has used a real-time method for salient object tracking on a mobile robotic platform. However, objects are learned here in a supervised manner with assistance of the operator.

Object of interest extraction is driven by its saliency, therefore accurate and fast salient region detection is crucial for our system. Although there exist numerous approaches described in the literature, not all of them are suitable for our purposes. Often they lack precision or good resolution in frequency domain, are only able to extract the one most salient object from the image, or are computationally too heavy to be used in real-time. A comparison of some state-of-the-art algorithms in these terms may be found in [2]. In this work, we propose a novel visual saliency allows to design a fast image segmentation algorithm, robust to real-world illumination conditions allowing to extract several objects from raw images. The visual saliency is used in the first part of our system. But before describing in detail the visual saliency in Sect. 3, we begin to give an overview of the system in the following sub-section.

2.2 System overview

The system we propose here consists of several units which collaborate together on the goal. On Fig. 1 a block-diagram of the system is depicted showing the individual units and their relations. Two main parts may be identified. The first one, labeled “Acquisition of new objects for learning” takes a raw image from the camera, detects visually important objects on it and extracts them so that they can be used as prospective samples for learning. These samples are then used in the second part (“Learning and recognition”), where learning of the extracted objects is done and thus further recognition of those objects is made possible.

Each one of the two mentioned parts contains several processing units. In the first unit, as a new image is acquired by the camera, it is processed by the “Salient region detection” unit (described in Sect. 3). Here, using hybrid features of chromaticity and luminosity along with local features of center-surround histogram calculation, a saliency map is constructed. It highlights regions of the image that are visually important, i.e. that are visually more salient with respect

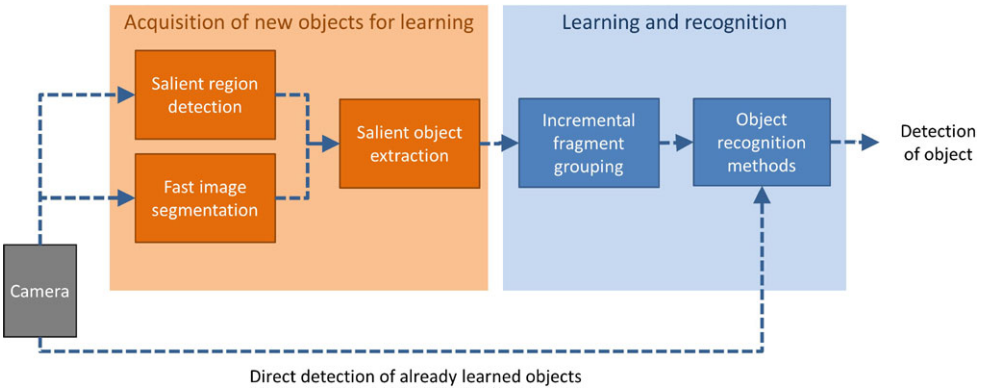

Fig. 1 Block diagram of our system and it's units

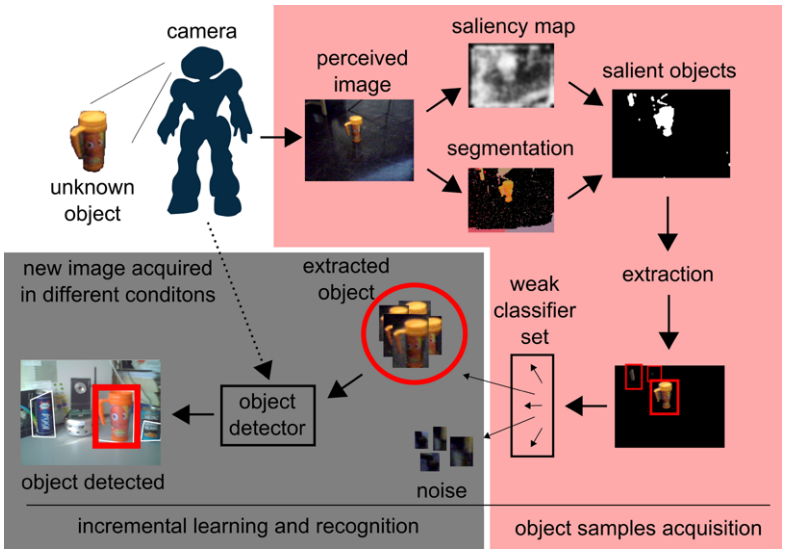

Fig. 2 Overview of the entire proposed system's work-flow. An unknown object is incrementally learned by extracting it and providing it as learning samples to the object detector (*solid arrows*). This enables recognition the object when encountered again (*the dotted arrow*)

to the rest of the image. In parallel the input image is processed in the "Fast image segmentation" unit, which splits the image into a set of segments according to the chromatic surface properties. The algorithm is shown to be robust to common illumination effects like shadows and reflections, which helps our system to cope with real illumination conditions. Finally the "Salient object extraction" unit (Sect. 4) combines results of the two previous, extracting the segments found on regions that exhibit significant saliency and forming them together to present at the end salient objects extracted from the input image.

As images are taken consecutively by the camera, salient objects extracted from each one are fed into the "Incremental fragment grouping" unit (Sect. 5). Here, an on-line classification is performed on each object by a set of weak classifiers and incrementally groups containing the same object extracted from different images are formed. These groups can be then used as a kind of visual memory of visual database describing each of the extracted objects. This alone could be enough for recognition of each of the objects, if it was ensured that each particular object will be found in the same visual context (i.e. in the context where the object is salient with respect to its surroundings) next time it is encountered by our system. This is clearly too restrictive for a system with a goal to recognize the once learned objects in any condition. That is why we add the last unit of our system, tagged "Object recognition methods". Its role is, by employing existing object recognition algorithms, to learn from the visual database build by "incremental fragment grouping" unit and to recognize those objects regardless to their saliency in new settings. Thus for once learned objects, they can be recognized directly on the input image, which is denoted by the very bottom arrow on the Fig. 1 labeled "Direct detection of already learned objects".

A different view on our system is presented on Fig. 2, where its work-flow is visualized. In short, based on the previous description, we can argue that the proposed cognitive system allows to design an artificial visual system with the following key capacities:

- autonomous extraction of salient objects from raw unlabeled camera images,
- learning of those objects and further recognition of the learned objects,

– robustness in order to recognize the learned objects in different conditions or visual contexts,
– embedded system and real time computing for robotic applications.

The next section describes, in detail, the new salient object detection algorithm which is used in this work.

3 Salient region detection

In this section, we propose a novel visual saliency detector composed of two independent parts, which can be computed in parallel. The first part captures saliency in terms of hybrid distribution of colors (i.e. a global saliency characteristic, Sect. 3.2). The second part calculates local characteristics of the image using a center-surround operation (Sect. 3.3). Their resulting saliency maps are merged together using a translation function, resulting in the final saliency map (Sect. 3.4).

In the proposed saliency computation algorithm, we are representing colors using a spherical interpretation of RGB color space (siRGB further on). This allows us to work with photometric invariants instead of pure RGB information. In this section, we begin to remember the main principles of siRGB (Sect. 3.1) but there are several works that explain in detail the siRGB color space and photometric invariants, see [25] and [33].

3.1 A spherical interpretation of RGB color space

Any image pixel's color corresponds to a point in the RGB color space $\{R_c, G_c, B_c\}$. The vector going from the origin up to this point can be represented using spherical coordinates $c = \{\theta_c, \phi_c, l_c\}$, where θ is zenithal angle, ϕ is azimuthal angle and l is the vector's magnitude (intensity). In RGB color space, chromaticity Ψ_c of a color point is represented by its normalized coordinates $r_c = \frac{R_c}{R_c+G_c+B_c}$, $g_c = \frac{G_c}{R_c+G_c+B_c}$, $b_c = \frac{B_c}{R_c+G_c+B_c}$, such that $r_c + g_c + b_c = 1$. That is, chromaticity corresponds to the projection on the chromatic plane Π_Ψ, defined by the collection of vertices of RGB cube $\{(1,0,0),(0,1,0),(0,0,1)\}$, along the line defined as $L_c = \{y = k \cdot \Psi_c; k \in \mathbb{R}\}$. In other words, all the points in line L_c have the same chromaticity Ψ_c, which is a 2D representation equivalent to one provided by the zenithal and azimuthal angle components of the spherical coordinate representation of the color point. Given an image $\Omega(x) = \{(R,G,B)_x; x \in \mathbb{N}^2\}$, where x refers to the pixel coordinates in the image grid domain, we denote the corresponding spherical representation as $\Omega(x) = \{(\theta,\phi,l)_x; x \in \mathbb{N}^2\}$, which allows us to use $(\theta,\phi)_x$ as the chromaticity representation of the pixel's color. For computational purposes, further on we normalize the angle θ and ϕ and the value l into a range from 0 to 255.

3.2 Global saliency features

For the first part, calculation of color saliency is done using two features: the intensity saliency (1) and the chromatic saliency (2). Here we define the saliency as Euclidean distance of intensity l (or azimuth ϕ and zenith θ respectively) of each pixel to the mean of the entire image. Index l stands for intensity channel of the image, $\Omega_{\mu l}$ is the average intensity of the channel, similarly for azimuth ϕ and zenith θ in (2). The term (x) denotes coordinates of a given pixel on the image.

$$M_l(x) = \left\| \Omega_{\mu l} - \Omega_l(x) \right\| \tag{1}$$

$$M_{\phi\theta}(x) = \sqrt{\left(\Omega_{\mu\phi} - \Omega_\phi(x)\right)^2 + \left(\Omega_{\mu\theta} - \Omega_\theta(x)\right)^2} \tag{2}$$

The composite color saliency map M is a hybrid result of combination of maps resulted from (1) and (2). Blending of the two saliency maps together is driven by a function of color saturation of each pixel. For this purpose, we define the color saturation C_c. It is calculated from RGB color model for each pixel as pseudo-norm given by $C_c = \max[R,G,B] - \min[R,G,B]$. When C_c is low (too dark or too bright areas of the image), more importance is given to intensity saliency (1). When C_c is high (vivid colors), chromatic saliency (2) is emphasized. As blending function we use the logistic sigmoid, so that the composite saliency map M is calculated following (3) where $C = 10(C_c - 0.5)$.

$$M(x) = \frac{1}{1-e^{-C}} M_{\phi\theta}(x) + \left(1 - \frac{1}{1+e^{-C}}\right) M_l(x) \tag{3}$$

Similar feature as we compute in (1) is used by [2]. However its authors use only a single distance for all three channels, mixing chromaticity and intensity value of pixels together, while our approach respects the color saturation, which allows us to treat separately chromatic and achromatic regions. This is particularly helpful in cases where both chromatic and achromatic objects are present on the image.

3.3 Local saliency features

For the local saliency features, the idea is to go through the entire image and to compare the content of a sliding window with its surroundings to determine, how similar the two are. If similarity is low, it may be a sign of a salient region. Let us have a sliding window P of size p, centered over pixel (x). Define a (center) histogram H_c of pixel intensities inside it. Then let us define a (surround) histogram H_s as histogram of intensities in a window Q surrounding P in a manner that the area of $(Q–P) = p^2$. The center-surround feature d is then given as over all histogram bins (i).

$$d(x) = \sum_i \left| \frac{H_c(i)}{H_c} - \frac{H_S(i)}{H_S} \right| \tag{4}$$

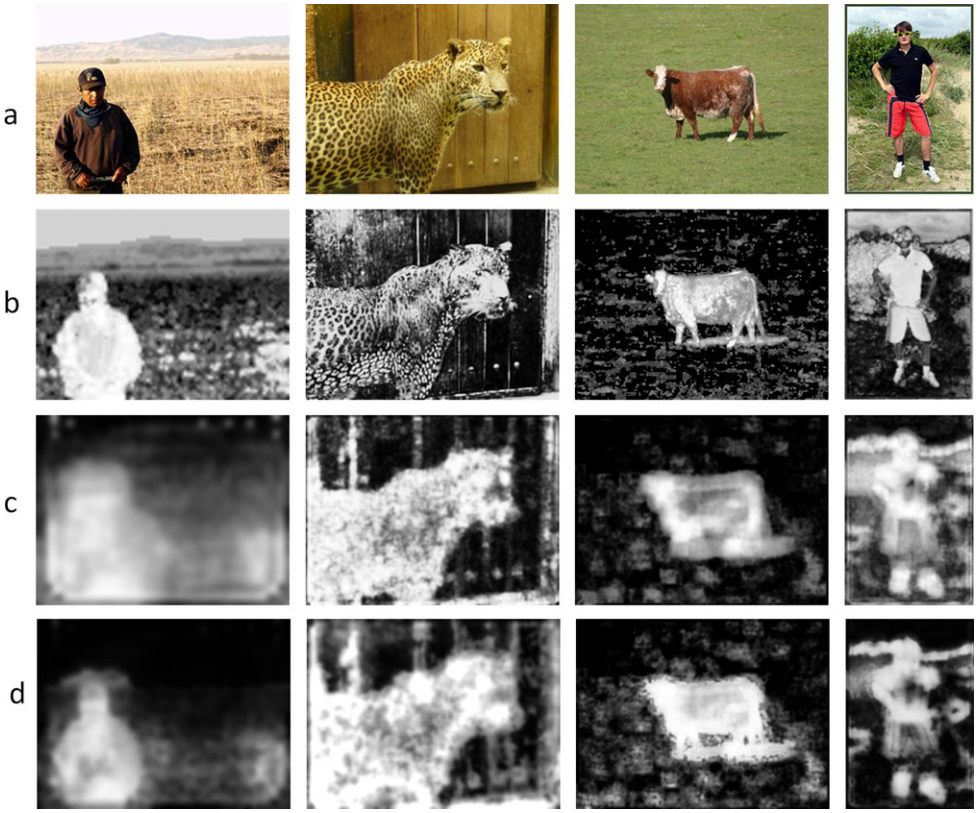

Fig. 3 **a**: Original images, **b**: composite saliency map M, **c**: center-surround saliency D, **d**: final saliency map M_{final}

Calculating the $d(x)$ (4) throughout all the l, ϕ and θ channels, we can compute the resulting center-surround saliency D on a given position (x) as follows in (5). To improve the performance of this feature on images with mixed achromatic and chromatic content, we use a similar approach of hybrid combination of chromaticity and intensity as we used in (3). However, here the color saturation C refers to average saturation over the sliding window P.

$$D(x) = \frac{1}{1-e^{-C}} d_l(x) + \left(1 - \frac{1}{1+e^{-C}}\right) \max\left(d_\phi(x), d_\theta(x)\right) \tag{5}$$

By using integral histograms described in [29], all the mentioned histogram operations can be done very efficiently in constant time with respect to parameter p. This parameter permits moreover a top-down control of the attention and of the sensitivity of the feature in scale space. High p value with respect to the image size will make the feature more sensitive to large objects; low values will allow more focus to smaller objects and details. Note however, that all the experiments described further on were carried out with a constant value of p fixed on 20 % of image width (i.e. 72 px for standard 320 × 240 px camera images).

3.4 Final saliency features

As the last step, both the global color saliency $M(x)$ from (3) and the local center-surround feature $D(x)$ from (5) are combined together by application of (6), resulting in final saliency map M_{final}, which is then smoothed by Gaussian filter. The upper part of the condition in (6) describes a particular case, where a part of the image consists of a color, that is not considered salient (i.e. pixels with low $M(x)$ measure), but which is distinct to the surroundings by virtue of its shape.

$$M_{final}(x) = \begin{cases} D(x) & \text{if } M(x) < D(x) \\ \sqrt{M(x)D(x)} & \text{else} \end{cases} \tag{6}$$

On Fig. 3 some resulting saliency maps of the presented algorithm are shown. Note that for the second image (leopard) the saliency map M (i.e. the global features, row b) does not highlight entirely the leopard's body. This image was selected to illustrate cases, where saliency consists in shape or texture of an object, which is distinct to its surroundings, rather than simply in its color. To capture this aspect of saliency, we compute the second (local) feature over the image: a center-surround difference of histograms (feature inspired by [21]). Regarding the features we use for saliency map calculation, our algorithm belongs to the group of saliency detection approaches, which are not able

to cope with psychological patterns like "curve", "intersection", "closure" etc. However, we do not perceive this as a shortcoming as our algorithm is primarily aimed for processing natural images and not to mimic precisely human psychological or vision system.

Having the saliency map of the input image computed, we can proceed to extraction of visually salient objects themselves. This is the purpose on the next section.

4 Salient object extraction

Manual fixed-value thresholding on the final saliency map and automatic thresholding using the Otsu's method have proven themselves as impracticable as well as other statistics based methods that we have applied on the saliency map. The problem is that all these methods work only over the saliency map and do not take into account the original image. Given this observation, we have decided to first apply a segmentation algorithm on the original image to obtain coherent parts of it; then we extract only those segments that are salient enough.

4.1 Main segmentation problems

There are sophisticated techniques for image segmentation like growing-neural-gas approaches applied in real time [4, 14], however we are going to focus in reflectance physics properties of the image for our segmentation process. Let us review some previous definitions about segmentation.

Image segmentation can be defined as a process which divides an image into different regions such that each region has a particular property, but the union of any two adjacent regions is not homogeneous. A formal definition of image segmentation is given in [13]. According to [13]: If $W()$ is a homogeneity predicate defined on groups of connected pixels, then segmentation is a partition of the set F into connected subsets or regions $(S_1, S_2, \ldots, S_n)$ such that with $\bigcup_{i=1}^{n} S_i = F$ and $\forall i \neq j,\ S_i \cap S_j = \emptyset$ and $\forall x, y \in S_i;\ W(x) = W(y)$.

There are four main problems in image segmentation: problems derived of the illumination, noise effects, edge ambiguity and the computational cost. These three first problems are closely related. In segmentation processes the use of a suitable distance measure is very important. Therefore we introduce a hybrid distance which works with intensity and chromaticity. On one hand, this hybrid distance allows parametrization of noise tolerance and on the other hand, we can adapt this distance for optimal edge detection. Furthermore, this distance is grounded in the dichromatic reflection model from [32] by a spherical interpretation of the RGB color space from [27]. So, this approach helps to avoid the first mentioned problem as well. Finally, in this method we will use only the 4-west-nord neighbors; it helps to decrease the computing time. The presented segmentation algorithm has thus the following properties: a good behavior in shadows and shines, avoids the effect of noise and finally it is cheap in terms of computing time.

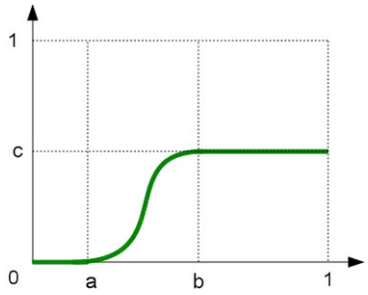

Fig. 4 Chromatic activation function $\alpha(x)$

4.2 Distance

We propose a distance based in the spherical interpretation of the RGB color space. Given an image $\Omega(x) = \{(R, G, B)_x; x \in \mathbb{N}^2\}$ where x refers to the pixel coordinates in the image grid domain, we denote the corresponding spherical representation as $\Omega(x) = \{(\phi, \theta, I)_x; x \in \mathbb{N}^2\}$, which allows us to use $(\phi, \theta)_x$ as the chromaticity representation of the pixel's color.

Empirical experiments tell us that intensity is the most important clue in dark regions, and that on the other hand it is better to use the chromaticity component when the illumination is good. Like in previous works [26] we propose a hybrid distance. Figure 4 shows the chromatic activation function. For values less than a, the chromatic component is inactive, for values that belong to the interval $[a, b]$, we take into account the chromatic component from its minimum energy to its maximum energy c by following a sinusoidal shape. Finally for values bigger than b its energy is always c. The three parameters a, b, c are in the range $[0, 1]$. The region under the green line is the chromatic importance and its complementary, the region over this line is the intensity importance. The function $\alpha(x)$ depends of the image intensity. Its complementary function $\bar{\alpha}(x)$ is the intensity activation function where $\bar{\alpha}(x) = 1 - \alpha(x)$ and hence $\bar{\alpha}(x) + \alpha(x) = 1$. The below equation is the mathematical expression of (x):

$$\alpha(x) = \begin{cases} 0 & x \leq a \\ \frac{c}{2} + \frac{c}{2}\sin(\frac{(x-a)\pi}{b-a} + \pi) & a < x < b \\ c & x \geq b \end{cases} \tag{7}$$

Now we can formulate a hybrid distance between any two pixels p, q as follows:

$$d_h(p, q) = \bar{\alpha}(p, q) \cdot d_I(p, q) + \alpha(p, q) \cdot d_\Psi(p, q) \tag{8}$$

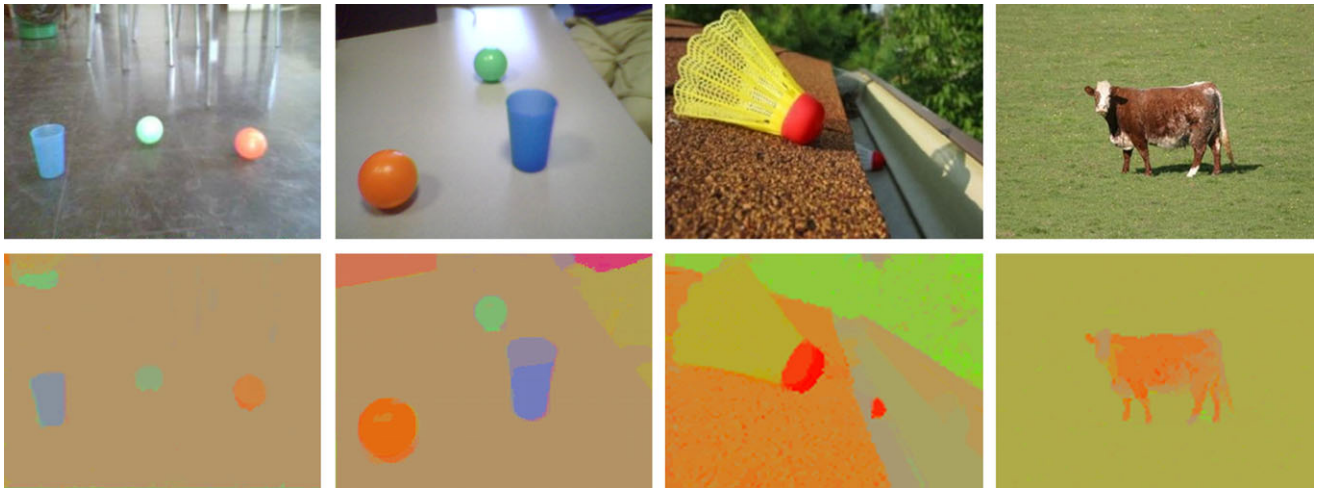

Fig. 5 Sample results of our segmentation algorithm. *Top row*: original images, *bottom row*: resulting segmentation

where the relationship between $\alpha(x)$ and $\alpha(p,q)$ is done by $x = \frac{l_p+l_q}{2}$ where l_p, l_q are the intensities l in spherical coordinates. d_l is an intensity distance as $d_l(p,q) = |l_p - l_q|$ and d_Ψ is a chromatic distance as $d_\Psi(p,q) = \sqrt{(\theta_q - \theta_p)^2 + (\phi_q - \phi_p)^2}$.

4.3 Segmentation algorithm

All of the previously described techniques are joined here covering the four desired goals. On one hand we are going to use the spherical interpretation of the RGB image, and on the other hand we are going to use the aforementioned hybrid distance expressed in Eq. (8). For edge detection a formal gradient is not necessary because it can be calculated "ad-hoc" using our hybrid distance and using a threshold. In fact this method is focused in the detection of homogeneous regions. When the distance between some pixels is less than this threshold we are going to admit that these pixels are homogeneous and then they belong to the same region, otherwise an edge is present.

In order to decrease the computing time, we use a 4-west-nord neighborhood, that is, by columns and rows, computing each pixel only one time. Homogeneous convex regions are easily identified because all of them have the same label. Our method is explained by the algorithm given in appendix. This algorithm returns a bi-dimensional integer matrix of labels. For the computation of this algorithm we also need a structure that relates each label with a chromaticity and the number of pixels labeled with it. That is necessary because each time we assign a new pixel to a label we must actualize the chromaticity of this label, which is the mean chromaticity of all pixels labeled with it. The most important parameter for this algorithm is the threshold δ. The granularity and noise tolerance depend on this. For a very little value we will obtain a lot of small regions, opposed to, with a high value we obtain the large and visually more important regions. On the other hand the parameters a, b, c of Eq. (8) allows to adjust the distance type. If $b = 0$ and $c = 1$ it is a pure chromatic distance. If $a = 1$ it is a pure intensity distance. In other cases it is a hybrid distance. In this algorithm $L(x)$ denotes the label of pixel x, $L_4(x)$ denotes the set of labels of the 4-west-north neighbors of pixel x, that can be expressed as $L_4(x) = \bigcup_{x' \in N_4(x)} L(x')$, where $N_4(x)$ the 4-west-north neighborhood of pixel x. The algorithm may be applied to any color image $\Omega(x)$. It needs the specification of the distance $d_H(x,y)$ that gives a measure of the similarity between pixel colors $\Omega(x)$ and $\Omega(y)$. To label the regions we keep a counter R, and we build a map Ψ_R assigning to each region label a chromatic value. We also have a counter C_R of the number of pixels in the image region of label R.

On Fig. 5, sample results of the described segmentation algorithm are shown. In particular, on first three images, the way how the algorithm reacts to difficult illumination conditions like highlights and shadows is presented. Correct segmentation results are obtained even in presence of strong light and reflections.

4.4 Extraction of salient objects using segmented image

The segmentation algorithm splits an image into a set of chromatically coherent regions. Objects present on the scene are composed of one or multiple such segments. For objects that conform to conditions of "explicitness", the segments forming them should cover areas of saliency map with high overall saliency, while visually unimportant objects and background should have this measure comparatively low.

Input image is thus segmented into connected subsets of pixels or segments $(S_1, S_2, \ldots, S_n)$. For each one of the found segments $S_i \in \{S_1, S_2, \ldots, S_n\}$ its average saliency $\overline{S_i}$

is computed over the saliency map M_{final} as well as the variance of saliency values $Var(S_i)$. All the pixel values $p(x, y) \in S_i$ of the segment are then set following (9), where $t_{\overline{S}}$ and t_{Var} are thresholds for average salience and its variance respectively. The result is a binary map containing a set of connected components $C = \{C_1, C_2, \ldots, C_n\}$ formed by adjacent segments S_i evaluated by (9) as 1. To get rid of noise, a membership condition is imposed that any $C_i \in C$ has its area larger than a given threshold. Finally, we project the binary map on the original image, which gives as a result parts of the original image containing its salient objects. For our experiments, we set $t_{\overline{S}}$ to 50 % of the maximal possible saliency, t_{Var} to 20 and the minimal area to 1 % of total image area.

$$\forall S_i \in \{S_1, S_2, \ldots, S_n\};\ \forall p(x, y) \in S_i;$$
$$p(x, y) = \begin{cases} 1 & \text{if } \overline{S_i} > t_{\overline{S}} \text{ and } Var(S_i) > t_{Var} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

5 Learning and recognition

The approach described in Sects. 3 and 4 allows us to split an image into a set of fragments, each containing a visually salient object. In this section we explain how we use this to enable a machine vision system to learn an object from unlabeled images. For experiments in real environment described further on we have used a mobile humanoid robot equipped with a color CMOS camera as a source of images. When acquired, images are processed to extract fragments containing salient objects; those fragments are grouped online using approach presented in Sect. 5.1. Only groups with a significant number of members are used as samples database for object recognition methods (see Sect. 5.2), which permits recognition of previously seen objects in different visual context or environment and moreover enables learning multiple objects in the same time.

5.1 Incremental fragment grouping

To preserve as much as possible the on-line and real time nature of learning, we have to group image fragments incrementally as they come from salient object detector with comparatively low calculation efforts. For this task we employ a combination of weak classifiers $\{w_1, w_2, \ldots, w_n\}$, each one classifying a fragment as belonging (result 1) or not belonging (result 0) to a certain class. Each classifier has a high level of false positives but a very low level of false negatives. In our case we employ four weak classifiers ($n = 4$), covering different properties of object on the fragment. A fragment belongs to a class if $\prod_{i=1}^{n} w_i = 1$. A class is allowed to be populated only once by one fragment per image to prevent overpopulation by repeating patterns on the same image. If a fragment is not put into any class by classifiers, a new class is created for it. If a fragment satisfies this equation for multiple classes, it is assigned to the one whose Euclidean distance is smaller in terms of features measured by each classifier (i.e. c_{wn}). Features taken into account by weak classifiers are as follows. In all equations, F denotes the currently processed fragment, whereas G denotes an instance of the group in question. All other symbols are explained further on in the text.

Area: the w_1 in Eq. (10) classifier separates fragments, whose difference of areas is too large. In experiments, we set t_{area} to 10.

$$w_1 = \begin{cases} 1 & \text{if } c_{w1} < t_{area} \\ 0 & \text{otherwise} \end{cases}$$
$$\text{where } c_{w1} = \frac{\max(G_{area}, F_{area})}{\min(G_{area}, F_{area})} \tag{10}$$

Aspect: the w_2 in Eq. (11) classifier separates fragments, whose aspect ratios are too different to belong to the same object. In experiments, we set t_{aspect} to 0.3.

$$w_2 = \begin{cases} 1 & \text{if } c_{w2} < t_{aspect} \\ 0 & \text{otherwise} \end{cases}$$
$$\text{where } c_{w2} = \left| \log\left(\frac{G_{width}}{G_{height}}\right) - \log\left(\frac{F_{width}}{F_{height}}\right) \right| \tag{11}$$

Chromaticity distribution: the w_3 in Eq. (12) classifier separates fragments with clearly different chromaticity. It works over 2D normalized histograms of ϕ and θ component of fragment denoted by $G_{\phi\theta}$ and $F_{\phi\theta}$ respectively with N histogram bins, calculating their intersection. We use N equal to 32 to avoid too sparse histogram and $t_{\phi\theta}$ equal to 0.35.

$$w_3 = \begin{cases} 1 & \text{if } c_{w3} < t_{\phi\theta} \\ 0 & \text{otherwise} \end{cases}$$
$$\text{where } c_{w3} = \frac{\sum_{j=1}^{N-1} \sum_{k=1}^{N-1} \min(G_{\phi\theta}(j,k) - F_{\phi\theta}(j,k))}{L^2} \tag{12}$$

Texture uniformity: the w_4 in Eq. (13) classifier separates fragments, whose texture is too different. We use the measure of texture uniformity calculated over the l channel of fragment. In (13), $p(z_i); i = 0, 1, 2, \ldots, L-1$ is a normalized histogram of l channel of the given fragment and N is the number of histogram bins. In experiments, we use 32 histogram bins to avoid too sparse histogram and value $t_{uniformity}$ of 0.02.

$$w_4 = \begin{cases} 1 & \text{if } c_{w4} < t_{uniformity} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } c_{w4} = \left| \sum_{j=0}^{N-1} p_G^2(z_j) - \sum_{k=0}^{N-1} p_F^2(z_k) \right| \quad (13)$$

5.2 Object recognition methods

Although we are able to extract individual objects by means of their visual saliency, this ability alone cannot be used for their further re-detection in different conditions e.g. by a mere comparison of the extracted object with the ones already acquired. It is because there is no guarantee that next time we encounter the object it will be distinct to its surroundings (i.e. salient) and it won't be cluttered or partially obscured by other objects. To cope with this, we use existing object recognition approaches to detect in new conditions the objects we already acquired.

In any time of learning the fragment grouping algorithm provides us a set of groups, each one populated by fragments of images containing the same objects. We can chose any of them and use fragments contained in it as a database of samples for an object recognition algorithm. For detection of objects in context of our system, any suitable real time recognition algorithm can be used. As a demonstration, we have employed two recognition algorithms. Here we will explain the basics of their function and how they make use of data about objects acquired in form of groups of fragments.

The first object recognition technique we used Speed-up Robust Features, or SURF, described in [6] is a well-established technique based on matching interest points on the source image with interest points coming from the template. It provides detection robust to partial occlusions and perspective deformations. In our case we use the fragments acquired as matching templates. To preserve the real-time operation of detection even with high numbers of templates, we pre-extract key points from each template in advance. In the detection stage, we first match several parallel threads templates with the greatest number of key-points (i.e. containing more visual information) and stop this process when another image from the camera arrives. This gives us an opportunity to test up to few tens template matches per frame. Further important speed-up can be achieved using parallel computation power of CUDA-like architectures on modern GPUs.

The second object recognition method, used in our work, is Viola-Jones detection framework published in [34]. Its principle is based on browsing sub-windows over the image and a cascade of classifiers, which determine whether the processed part of image belongs or not to a group of objects on which the classifier was trained. In this case, we use acquired fragments of an object as positive samples to learn the cascade of classifiers (the learning here is carried out offline due to the nature of this method). As this method requires negative samples as well, we use the original images with the learned object replaced by a black rectangle. To be precise enough, the method needs up to several thousands of samples for learning. We achieved this number by applying random contrast changes, perspective deformations and rotation of learning fragments. Although Viola-Jones framework was originally designed to recognize a class of objects (i.e. human faces), rather than single instances, in our case we use it in the way that it recognizes a class of only one object (i.e. the one found on learning fragments). It must be noticed that having the object detector learned, known objects can be detected directly from the input image when seen again, without passing by the salient object detection.

6 Results

As we have written in the introduction of this paper, the goal of this work was to design an intelligent machine vision system able to learn autonomously individual objects present in a real environment. This section is dedicated to present the main obtained results. In first, we begin to present intermediate results about the salient object detection. Although we did not do a comparative study between state of the art saliency methods (e.g. [2] and [21]) and ours, we give some points of comparison. More results and explanations are given in [30]. In Sect. 6.2, we give results obtained on a real robot. Finally, we discuss these results.

6.1 Salient object detection results

On Fig. 6, sample results of our algorithm are compared with ground truth and two others state of the art algorithms. It must be pointed out that our initial goal was not to design a new saliency method but rather to adapt existing approaches in the frame work of robotic applications. These applications imply , especially, robustness and real-time computing. But in order to evaluate the performance of the proposed solution, we have chosen to compare our work with the work presented in [2] and [21] because the first one presents a fast algorithm potentially suitable for real-time application in machine vision, while the latter one shows high performance in terms of precision and correctness. No claims are made by authors of the latter one about its speed, but with respect to the description of the algorithm provided in [21] we assume that it is not suitable for a real-time application.

Figure 6 show results which illustrate the typical performance of presented algorithms. Although [2] is computationally very cheap (saliency map calculation takes about 45 ms on a 320 × 240 px image), its results vary largely in quality depending on the nature of salient objects on the image. Algorithm of [21] produces results very close to human perception and more precise in terms of resolution (sample results are published online[1]). However, it suffers from

[1] http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm.

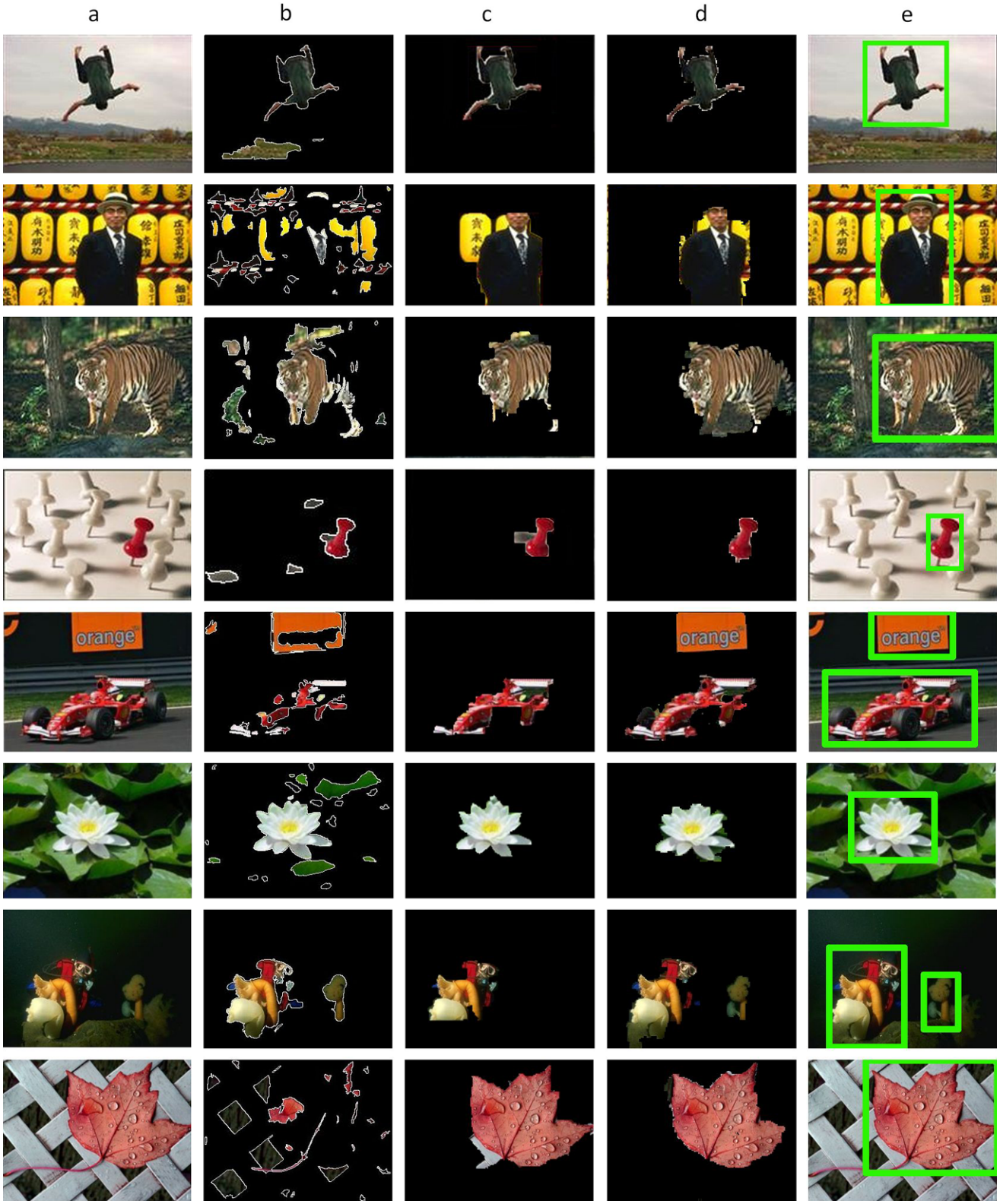

Fig. 6 Comparison of different salient object detection algorithms. **a** *First column*: original image. **b** *Second column*: results of (AC) [2]. **c** *Third column*: results of (LI) [21]. **d** *Fourth column*: results of our approach. **e** *Last column*: ground truth (taking into account multiple objects in the scene).

Table 1 Scores obtained by our salient object detection algorithm on the MRSA dataset

	Precision	Recall	F-measure
Dataset A	0.73	0.75	0.74
Dataset B	0.75	0.76	0.75

two major drawbacks in context of the learning system we present here. It does not claim to be applicable in real time, and more importantly it outputs only one salient objects (i.e. the most salient object) a time (although authors suggest for future work a workaround to this using inhibition-of-return technique). On the other hand our approach outputs natively multiple salient objects if they are present on the image. An illustrative example may be found on Fig. 6, the fifth row, where two visually attractive objects are found on the same image: the F1 racing car and the "orange" logo. As they are both highly salient and clearly distinct in terms of their position on the image, our algorithm marks them both as visually salient. This property appears to be crucial while extracting unknown objects for learning as there is no reason why only the most salient object should be learned, especially in real conditions with highly structured environment and many objects present in the field of view.

We tested our algorithm against the benchmark on MSRA Salient Object Database [21] by using the same protocol. The results are given in Table 1. While these results are close to results obtained by [21] (the F-measure differs from the [21] only by about 0.05), our algorithm brings the benefit of high-speed processing and native output of multiple salient regions, if they are present on the image. In terms of average speed, on 320 × 240 px the method of AC calculated the saliency map in 45 ms, but takes another 2900 ms per image to extract salient segments using mean-shift segmentation.[2] Our algorithm in its unoptimized version takes in average 100 ms per image (saliency map and image segmentation are calculated in parallel as they are independent processes), which allows us to run it at a speed of about 10 frames per second. All algorithms were run on an Intel i5 CPU at 2.25 Ghz machine.

On Fig. 7 performance of our algorithm in some particular cases is illustrated:

- The case a shows how the algorithm copes with difficult illumination conditions in presence of strong directional light. The ball is extracted correctly and the strong reflection is not marked as a salient object as the shine does not change the chromatic property of the surface.
- In case of b and c, we can observe changes in parameter p of (5). In the first case (high p), the saliency is focused towards larger objects (extracting mainly the human head), whereas in the second case (small p) the emphasis is put on details, which allows to extract eyes, mouth, hair and other small details on the image.
- Finally, cases d and e capture extraction of objects with color and texture close to image background.

6.2 Results of validation a robots' vision

To verify the performance of our system, we have conducted a number of experiments with learning objects present in a common office environment. For the sake of repeatability and convenience in evaluation of results, we have collected ten common house or office objects to be explicitly learned (although the system naturally learns salient objects in its

[2]Based on executable available on http://ivrg.epfl.ch/supplementary_material/RK_CVPR09/index.html.

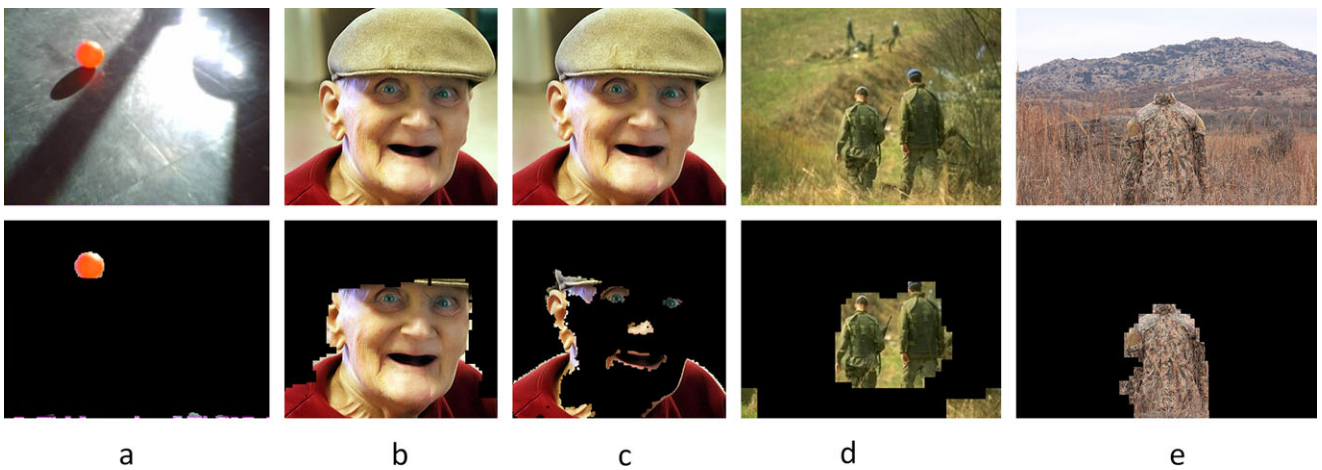

Fig. 7 Particular cases of our algorithm; **a**: conditions of strong illumination with reflections and shadows; **b**: human face with saliency detector focused on large objects; **c**: the same scene with saliency detector focused on small details; **d–e**: objects with camouflage close to the background (soldiers and a military combat uniform)

Fig. 8 Images of objects used throughout described experiments

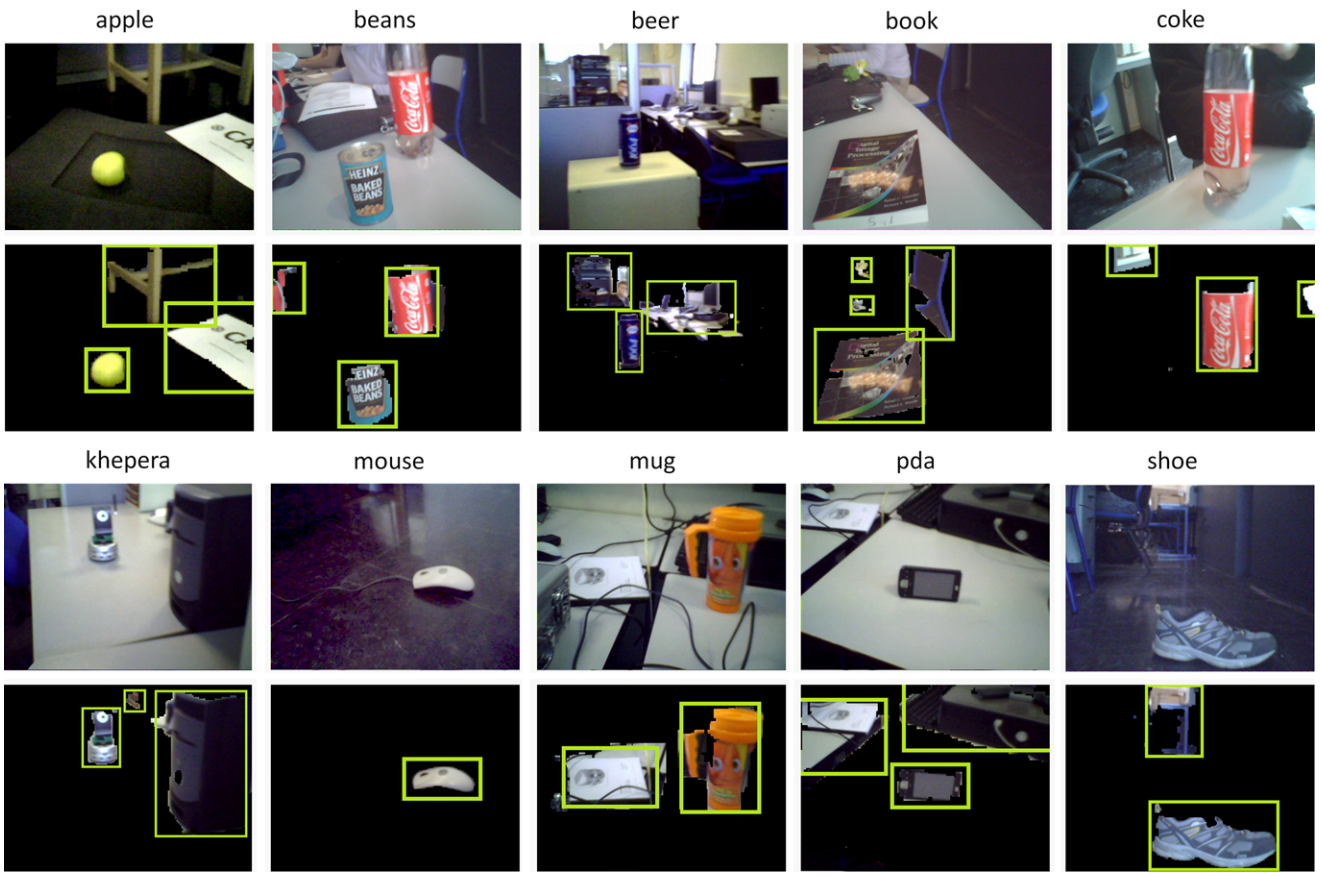

Fig. 9 Sample images from the training sequence for each object. Fragments containing salient objects detected by our algorithm are marked by *rectangles*

surroundings without a specific preference). Illustrative photos of the objects we used in this experiment are shown on Fig. 8 to give the reader a better idea about their nature and a sample of scene images containing those objects is presented of Fig. 9. To approach to the real conditions as much as possible we have chosen objects with different surface properties (chromatic, achromatic, textured, smooth, reflective, ...) and put them in a wide variety of light conditions and visual contexts. The number of images acquired for scenes containing each object varied between 100 and 600 for learning image sequences and between 50 and 300 for testing sequences, with multiple objects occurring on the same scene. Note that the high number of learning images were taken primarily in order to test our saliency detection and segment grouping algorithm. The learning process itself would generally require significantly less samples ac-

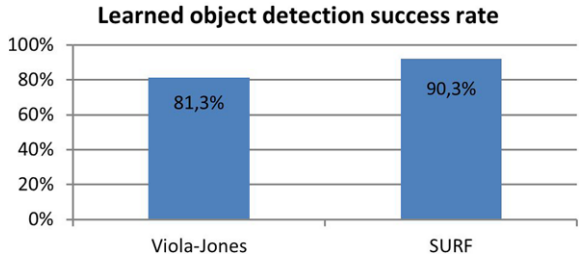

Fig. 10 Percentage of correct detections of learned objects over testing image set using Viola-Jones algorithm and SURF algorithm

quired to perform sufficiently well, depending on the actual object detection algorithm employed. On Fig. 9 random images from the learning sequence are presented for each learned object along with fragments extracted from them. These fragments, containing salient objects found on each image are subsequently processed by the incremental fragment grouping and the selected ones are used for learning of the object detector.

First we present results of salient object extraction and fragment grouping. To investigate the effectiveness of the salient object extraction, we have counted the percentage of learning set images, on which the learned object have been correctly detected and extracted by the salient object detector. Correct extraction means that the object has been extracted entirely and without any other object co-occurring on the fragment. We have achieved successful extraction on 82 % of images in the set. The subsequent grouping of fragments has achieved on the same data-set success rate of 96 %, i.e. only 4 % of fragments (usually bearing visual resemblance to other objects on the scene) were placed into a wrong group, which contained a fragments from a different object. On Fig. 10 detection rates over testing data-set using a trained Viola-Jones detection framework are provided along with performance of SURF algorithm. In average over all the objects in the testing set, the detection rate for Viola-Jones was about 81.3 %. In case of SURF, the average detection rate was higher at about 90.3 %. The numbers reflect only true positive detections. Average rate of false positive detections was insignificant (around 0.5 % in both cases).

To demonstrate real-time abilities of our system, we have successfully run several experiments, where a mobile robot equipped with color camera was required to learn a presented object. When learned, the robot was required to find the object in its environment and to track and follow it. Images from a video[3] acquired during those experiments are shown on Fig. 11. On Fig. 12, sample detection results are shown with a system having learned several objects. Boundary lines determine objects previously encountered by the robot and successfully recognized on the new scene.

6.3 Discussion

In Sect. 6.1, some qualitative results of our salient object extraction technique were given and compared with existing approaches. We have seen that the quality of our approach is comparable with the existing ones, while bringing advantage of real-time processing, native extraction of multiple salient objects and robustness to certain difficult illumination conditions. This confirms that your salient object extraction technique is performing enough to play its part in the proposed system. On Fig. 9, some qualitative results of our salient object extraction in real environment are given. They show typical views acquired during the learning of the system. Salient objects extracted from them. On all images (except of the "mouse" one, where only one object is present) multiple visually important objects were extracted apart of the one we placed intentionally to the scene, which is a desired behavior as the system is expected to extract (and learn) autonomously the encountered objects without any a-priory preference. On the other hand, as illustrated on the "mouse" and "shoe" images, the algorithm does not extract the "false objects" created by reflections found on the floor. The percentage of successfully extracted samples of a same object usable to learn this object is 82 % of its occurrences throughout a sequence of images. This means that roughly 4 of 5 acquired images of that object contribute in fact to correct learning of this object. As our learning system is incremental, the needed number of sample images for each object can be achieved accurately and fast enough.

We have employed two fundamentally different object recognition algorithms in our system, each yielding a different rate of recognition. The SURF has shown superior performance 90.3 % of average detection rate by contrast to Viola-Jones framework, which performed 9 % worse. This shows that Viola-Jones framework may not be best suited for this kind of task. We presume that it is mainly because of the fact that in order to achieve high recognition rates it typically needs thousands of learning samples, however the number of unique samples acquired for each learned object was in order of hundreds. Also its long learning time makes it impractical in the strict sense of on-line learning. On the other hand results achieved with SURF are encouraging both for the relatively high percentage of correct recognitions and for the fact that it allows recognition of the learned object even with only several samples acquired. Some camera views of

[3] This video can be found to the following address http://www.youtube.com/watch?v=xxz3wm3L1pE.

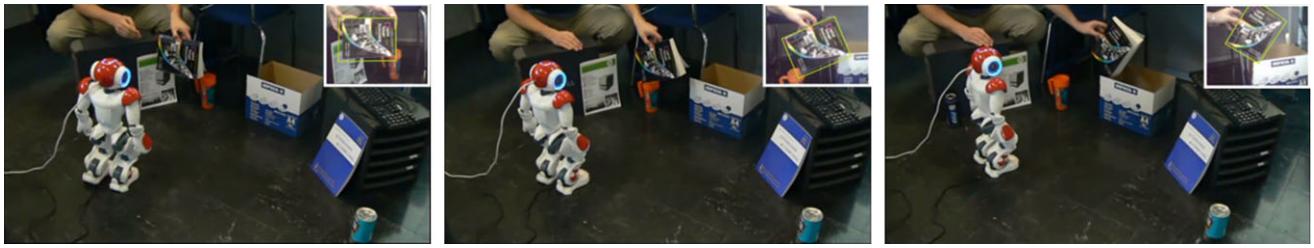

Fig. 11 Images from tracking a previously learned moving object. Robot camera picture is shown in *upper right corner* of *each image*

Fig. 12 Camera pictures from single (*first row*) object detection and multiple (*second row*) previously learned object detection

already learned system are shown on Fig. 12 with recognized objects marked by bounding shapes. On the first row individual objects are correctly detected. On the second row, three views on similar scenes containing multiple objects is shown. Between the scenes the system was progressively learning new objects so that e.g. the orange mug is recognized only on the last scene. The images show flexibility of recognition of the learned objects that are recognized in different orientation and perspective (the book), different illumination conditions (the shoe) or different distance and orientation (the coke bottle).

Regarding experiments with a robot searching for or tracking a previously learned object (Fig. 11), our system was successfully validated. It has enabled the robot to fulfill the required tasks, correctly responding to the input. Because of limited computing capacity of the robot used, we have chosen to run the system on a remote computer. In this experimental context, despite of the specific communication protocol implemented (by the constructor) on the robot, our system itself has been capable of real-time processing performance. However, one may observe a slow-down in robots reactions due to the limited bandwidth. This is the consequence of inadequacy of the aforementioned protocol regarding image transfer.

We have also identified certain shortcomings in learning chain naturally bound to the method of object extraction we are using. In fact, our system shows worsening performance in learning objects that are not visually distinct enough with respect to their background. The same happens in cases where two visually important objects are seen one behind another and thus are wrongly extracted as one by our current system. In effect, in order to respond correctly to this complex situation an additional level of machine intelligence would be necessary. However, it is pertinent to emphasize that in the actual system, once an object is correctly learned, its further detection (thanks to the object detectors employed) is practically independent from its visual context.

7 Conclusion and further work

In this work we have proposed an intelligent machine learning system with capacity of autonomous learning of objects

present in real environment. In its conception we were inspired by early processing stages of human visual system. In this context we suggested a novel algorithm for visually salient object detection, taking advantage of using photometric invariants. The algorithm has low complexity and can be run in real-time on contemporary processors. Moreover it exhibits robustness to difficult real-world light conditions. This algorithm is the first key part of the proposed machine vision system. We demonstrate that the detected salient objects can be efficiently used for training the second key part of our system, which is a machine learning-based object detection and recognition unit. Encouraging results were obtained especially when SURF detector was employed as an object detector.

In the future our approach could evolve in several ways. As there does not exist a universal recognition algorithm that suits any existing class of objects, other object detection algorithms, like GLOH in [24] or receptive field co-occurrence histograms in [11], could be adopted along with surface descriptors for each learned object. Objects of different characteristics could be then learned by algorithms that best suit the nature of the object in a "mixture of experts" manner. As to our visual saliency detector, the center-surround feature detector could be supplied by or replaced by an interesting approach of spectral residua detection published in [17]. A top-down feedback based on already acquired and grouped fragments could also greatly improve the saliency detector. The results presented here have been achieved with monocular camera. However, there are reasons to believe that the performance of our system could be enhanced by use of a stereo camera, where the depth-separation of objects would serve side-by-side with the segmentation algorithm to cope with the mentioned cases, where two visually important objects are one behind another.

An open question is, whether the presented technique, instead of learning solely individual objects, could be used as well for place learning and recognition, extracting visually important objects from the entire place like room or office or for visual navigation of a mobile robot. It would also be interesting to investigate, whether the saliency-based method could have an overlap outside the image processing domain, to be applied for learning of other than visual data (e.g. audio).

Appendix: Image segmentation algorithm in siRGB

Input: $\Omega(x)$ //the color image in spherical coordinates
$\delta,\ a,\ b,\ c$ //threshold and three distance parameters values
Output: bi-dimensional matrix L containing pixel labels l and an array of region chromaticity representations $\Psi_l\ \forall\ l$

$x_0 = \Omega(0,0) : L(x_0) = newlabel,\ \Psi_{l(x_0)} = \Psi(x_0)$
for *each* x **do**
  if $L_4(x) = \{l\}$ **then**
    //there is only one label in $N_4(x)$
    evaluate_neighbor (x)
  else
    $D \leftarrow \{d_h(L(y), L(z)) = d_{y,z} | y, z \in N_4(x)\ \&\ L(y) \neq L(z)\}$
    for *all* $d_{y,z} \in D$ ***s.t.*** $d_{y,z} < \delta$ **do**
      //region merging
      $L(y) = merge(y, z)$ //merge both regions into
      $L(y)\ \Psi_{L(y)} = avg\,\Psi(w), \forall w \in \Omega$ where
      $L(w) = L(y)$ //update reg. chroma.
    end
    if $L_4(x) = \{l\}$ **then**
      //there is only one label in $N_4(x)$
      evaluate_neighbor (x)
    else
      $d \leftarrow \min\{d_h(x, y) | y \in N_4(x)\}$
      if $d < \delta$ **then**
        //assign to region with lower distance
        $L(x) = L(y)$ **s.t.** $d_h(x, y) = d$
        $\Psi_{L(y)} = avg\,\Psi(w), \forall w \in \Omega$ where
        $L(w) = L(y)$ //update reg. chroma.
      else
        //current pixel cannot be assigned to any region
        create_new_label(x)
      end
    end
  end
end

function: create_new_label(x)
  $L(x) =$ newlabel //create a new region label $\Psi_{L(x)} = \Psi(x)$

function: evaluate_neighbor(x)
  $d \leftarrow \min\{d_h(x, y) | y \in N_4(x)\}$
if $d < \delta$ **then**
  //neighbor colors are similar
  $\Psi_{L(y)} = avg\,\Psi(w), \forall w \in \Omega$ where $L(w) = L(y)$
  //update reg. chroma.
else
  create_new_label(x)
end

References

1. Achanta R, Estrada F, Wils P, Süsstrunk S (2008) Salient region detection and segmentation. In: International conference on computer vision systems (ICVS'08). Lecture notes in computer science, vol 5008. Springer, Berlin, pp 66–75

2. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: IEEE international conference on computer vision and pattern recognition (CVPR)
3. An SY, Kang JG, Choi WS, Oh SY (2011) A neural network based retrainable framework for robust object recognition with application to mobile robotics. Appl Intell 35:190–210. doi:10.1007/s10489-010-0212-9
4. Angelopoulou A, Psarrou A, Garcia Rodriguez J, Gupta G (2008) Active-gng: model acquisition and tracking in cluttered backgrounds. In: Proceeding of the 1st ACM workshop on vision networks for behavior analysis, VNBA'08. ACM, New York, pp 17–22
5. Avidan S, Shamir A (2007) Seam carving for content-aware image resizing. ACM Trans Graph 26
6. Bay H, Tuytelaars T, Gool LJV (2006) Surf: speeded up robust features. In: Leonardis A, Bischof H, Pinz A (eds) ECCV (1). Lecture notes in computer science, vol 3951. Springer, Berlin, pp 404–417
7. Bay H, Ess A, Tuytelaars T, Gool LV (2008) Speeded-up robust features (surf). Comput Vis Image Underst 110:346–359
8. Borba GB, Gamba HR, Marques O, Mayron LM (2006) An unsupervised method for clustering images based on their salient regions of interest. In: Proceedings of the 14th annual ACM international conference on multimedia, MULTIMEDIA'06. ACM, New York, pp 145–148
9. Bülthoff HH, Wallraven C, Giese MA (2008) Perceptual robotics. In: Siciliano B, Khatib O (eds) Springer handbook of robotics. Springer, Berlin, pp 1481–1498
10. Chen LQ, Xie X, Fan X, Ma WY, Zhang HJ, Zhou HQ (2003) A visual attention model for adapting images on small displays. Multimed Syst 9(4):353–364
11. Ekvall S, Kragic D (2005) Receptive field cooccurrence histograms for object detection. In: 2005 IEEE/RSJ international conference on intelligent robots and systems (IROS 2005), pp 84–89
12. Frintrop S, Kessel M (2009) Most salient region tracking. In: Proceedings of the 2009 IEEE international conference on robotics and automation, ICRA'09. IEEE Press, Piscataway, pp 758–763
13. Fu K, Mui J (1981) A survey on image segmentation. Pattern Recognit 13(1):3–16
14. García-Rodríguez J, García-Chamizo JM (2011) Surveillance and human-computer interaction applications of self-growing models. Appl Soft Comput (in press, corrected proof)
15. Harel J, Koch C, Perona P (2007) Graph-based visual saliency. In: Advances in neural information processing systems, vol 19, pp 545–552
16. Hossain M, Dewan M, Chae O (2012) A flexible edge matching technique for object detection in dynamic environment. Appl Intell 36:638–648. doi:10.1007/s10489-011-0281-4
17. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. IEEE Conf Comput Vis Pattern Recognit 2(800):1–8
18. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20:1254–1259
19. Kursun O, Favorov OV (2010) Feature selection and extraction using an unsupervised biologically-suggested approximation to Gebelein's maximal correlation. Int J Pattern Recognit Artif Intell 24(3):337–358. http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=51857641&lang=fr&site=ehost-live
20. Liang Z, Chi Z, Fu H, Feng D (2012) Salient object detection using content-sensitive hypergraph representation and partitioning. Pattern Recognit 45(11):3886–3901
21. Liu T, Yuan Z, Sun J, Wang J, Zheng N, Tang X, Shum HY (2011) Learning to detect a salient object. IEEE Trans Pattern Anal Mach Intell 33(2):353–367
22. Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the international conference on computer vision, Washington, pp 1150–1157
23. Meger D, Muja M, Helmer S, Gupta A, Gamroth C, Hoffman T, Baumann MA, Southey T, Fazli P, Wohlkinger W, Viswanathan P, Little JJ, Lowe DG, Orwell J (2010) Curious george: an integrated visual search platform. In: CRV. IEEE Press, New York, pp 107–114
24. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630
25. Mileva Y, Bruhn A, Weickert J (2007) Illumination-robust variational optical flow with photometric invariants. In:Hamprecht FA, Schnörr C, Jähne B (eds) DAGM-symposium. Lecture notes in computer science, vol 4713. Springer, Berlin, pp 152–162
26. Moreno R, Graña M, Zulueta E (2010) Rgb colour gradient following colour constancy preservation. Electron Lett 46(13):908–910
27. Moreno R, Graña M, d'Anjou A (2011) Illumination source chromaticity estimation based on spherical coordinates in rgb. Electron Lett 47(1):28–30
28. Navalpakkam V, Itti L (2006) An integrated model of top-down and bottom-up attention for optimizing detection speed. In: Proceedings of the 2006 IEEE computer society conference on computer vision and pattern recognition, CVPR'06, vol 2. IEEE Computer Society, Washington, pp 2049–2056
29. Porikli F (2005) Integral histogram: a fast way to extract histograms in Cartesian spaces. In: IEEE computer society conference on computer vision and pattern recognition, CVPR 2005, vol 1. IEEE Computer Society, Los Alamitos, pp 829–836
30. Ramik D, Sabourin C, Madani K (2011) Hybrid salient object extraction approach with automatic estimation of visual attention scale. In: 2011 seventh international conference on signal-image technology and Internet-based systems (SITIS), pp 438–445. doi:10.1109/SITIS.2011.31
31. Rutishauser U, Walther D, Koch C, Perona P (2004) Is bottom-up attention useful for object recognition? In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, vol 2. IEEE Press, Washington, pp 37–44
32. Shafer SA (1985) Using color to separate reflection components. Color Res Appl 10(4):210–218
33. van de Weijer J, Gevers T (2004) Robust optical flow from photometric invariants. In: ICIP, pp 1835–1838
34. Viola PA, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
35. Wang Y, Qi Y (2013) Memory-based cognitive modeling for robust object extraction and tracking. Appl Intell 1–16. doi:10.1007/s10489-013-0437-5

Dominik Maximilián Ramík received his Master of Information Science degree from University of Ostrava, Czech Republic in 2008. His Master degree research concerned use of modular artificial neural networks for human face recognition. He received his PhD in 2012 in signal and image processing from University of University Paris-Est Creteil (UPEC). His current research topic concerns processing of complex images using bio-inspired artificial intelligence approaches and consequent extraction of semantic information with use in mobile robotics control and industrial processes supervision.

Christophe Sabourin graduated in Electrical Engineering in 1992 from University of Poitiers, France and received his MSc in Automation and Computer Science from this same University in 1993. He received his PhD in Robotics and Control from University of Orleans, France in November 2004. In September 2005, he joined Senart-Fontainebleau Institute of Technology of University Paris-Est/Paris 12 where he works as Associate Professor in the Electrical Engineering Department. Since 2005, he has been a researcher and a staff member of SCTIC Research Division, one of the two research components of Images, Signals and Intelligent Systems Laboratory (LISSI/EA 3956) of University Paris-Est Creteil. His current interests relate to areas of complex and bio-inspired intelligent artificial systems, cognitive robotics, humanoid robots, collective and social robotics.

Ramon Moreno obtained his Doctor-Engineer degree in Computer Science from Basque Country University (Universidad del País Vasco) in 2012, Ramon Moreno is a post-doctoral researcher at Computational Intelligence Group of this university. His research interests relate Computer Vision, Hyperspectral imagery, Reflectance models, Evolutive Algorithms and Computational Intelligence.

Kurosh Madani received his PhD degree in Electrical Engineering and Computer Sciences from University Paris XI, Orsay, France, in 1990. From 1989 to 1990, he worked as Assistant Professor at Institute of Fundamental Electronics of Paris XI University, Orsay, France. In 1990, he joined Creteil-Senart Institute of Technology of University Paris-Est Creteil (UPEC), Lieusaint, France, where he worked from 1990 to 1998 as Assistant Professor. In 1995, he received the DHDR Doctor Habilitate degree (senior research Dr Hab. degree) from UPEC. Since 1998 he has worked as Chair Professor in Electrical Engineering of Senart Institute of Technology of UPEC. From 1992 to 2004 he was Head of Intelligence in the Instrumentation and Systems Laboratory (I2/JE 2353). Co-initiator in 2005 of the Images, Signals and Intelligent Systems Laboratory (LISSI/EA 3956), he is head of one of the two research groups of LISSI. He has worked on both digital and analog implementation of processors arrays for image processing, electro-optical random number generation, and both analog and digital ANN implementation. His current research interests include large ANN structures modeling and implementation, hybrid neural based information processing systems and their software and hardware implementations, design and implementation of real-time neuro-control and neural based fault detection and diagnosis systems. In 1996 he became a permanent member (elected Academician) of the International Informatization Academy and in 1997 was elected as Academician of the International Academy of Technological Cybernetics.