Hindawi Publishing Corporation
Applied Computational Intelligence and Soft Computing
Volume 2012, Article ID 354785, 11 pages
doi:10.1155/2012/354785

Research Article

Multilevel Cognitive Machine-Learning-Based Concept for Artificial Awareness: Application to Humanoid Robot Awareness Using Visual Saliency

Kurosh Madani, Dominik M. Ramik, and Cristophe Sabourin

Images, Signals and Intelligence Systems Laboratory (LISSI/EA 3956) and Senart-FB Institute of Technology, University Paris-EST Créteil (UPEC), Bât.A, avenue Pierre Point, 77127 Lieusaint, France

Correspondence should be addressed to Kurosh Madani, madani@u-pec.fr

Received 11 March 2012; Revised 12 May 2012; Accepted 20 May 2012

Academic Editor: Qiangfu Zhao

As part of "intelligence," the "awareness" is the state or ability to perceive, feel, or be mindful of events, objects, or sensory patterns: in other words, to be conscious of the surrounding environment and its interactions. Inspired by early-ages human skills developments and especially by early-ages awareness maturation, the present paper accosts the robots intelligence from a different slant directing the attention to combining both "cognitive" and "perceptual" abilities. Within such a slant, the machine (robot) shrewdness is constructed on the basis of a multilevel cognitive concept attempting to handle complex artificial behaviors. The intended complex behavior is the autonomous discovering of objects by robot exploring an unknown environment: in other words, proffering the robot autonomy and awareness in and about unknown backdrop.

1. Introduction and Problem Stating

The term "cognition" refers to the ability for the processing of information applying knowledge. If the word "cognition" has been and continues to be used within quite a large number of different contexts, in the field of computer science, it often intends artificial intellectual activities and processes relating "machine learning" and accomplishment of knowledge-based "intelligent" artificial functions. However, the cognitive process of "knowledge construction" (and in more general way "intelligence") requires "awareness" about the surrounding environment and, thus, the ability to perceive information from it in order to interact with the surrounding milieu. So, if "cognition" and "perception" remain inseparable ingredients toward machines intelligence and thus toward machines (robots', etc.) autonomy, the "awareness" skill is a key spot in reaching the above-mentioned autonomy.

Concerning most of the works relating modern robotics, and especially humanoid robots, it is pertinent to note that they either have concerned the design of controllers controlling different devices of such machines [1, 2] or have focused the navigation aspects of such robots [3–5]. In the same way, the major part of the work dealing with human-like, or in more general terms intelligent, behavior, has connected abstract tasks, as those relating reasoning inference, interactive deduction mechanisms, and so forth. [6–10]. Inspired by early-ages human skills developments [11–15] and especially human early-ages walking [16–19], the present work accosts the robots intelligence from a different slant directing the attention to emergence of "machine awareness" from both "cognitive" and "perceptual" traits. It is important to note that neither the presented work nor its related issues (concepts, architectures, techniques, or algorithms) pretend being "artificial versions" of the complex natural (e.g., biological, psychological, etc.) mechanisms discovered, pointed out, or described by the above-referenced authors or by numerous other scientists working within the aforementioned areas whose works are not referenced in this paper. In [20] Andersen wrote concerning artificial neural networks: "*It is not absolutely necessary to believe that neural network models have anything to do with the nervous*

system, but it helps. Because, if they do, we are able to use a large body of ideas, experiments, and facts from cognitive science and neuroscience to design, construct, and test networks. Otherwise, we would have to suggest functions and mechanism for intelligent behavior without any examples of successful operation." In the same way, those natural mechanisms help us to look for plausible analogies between our down-to-earth models and those complex cognitive mechanisms.

Combining cognitive and perceptual abilities, the machine (robot) shrewdness is constructed on the basis of two kinds of functions: "unconscious cognitive functions" (UCFs) and "conscious cognitive functions" (CCFs). We identify UCFs as activities belonging to the "instinctive" cognition level handling reflexive abilities. Beside this, we distinguish CCFs as functions belonging to the "intentional" cognition level handling thought-out abilities. The two above-mentioned kinds of functions have been used as basis of a multilevel cognitive concept attempting to handle complex artificial behaviors [21]. The intended complex behavior is the autonomous discovering of objects by robot exploring an unknown environment. The present paper will not itemize the motion-related aspect that has been widely presented, analyzed, discussed and validated (on different examples) in [21]. It will focus on perceptual skill and awareness emergence. Regarding perceptual skill, it is developed on the basis of artificial vision and "salient" object detection. The paper will center this foremost skill and show how the intentional cognitive level of the above-mentioned concept could be used to proffer a kind of artificial awareness skill. The concept has been applied in design of "motion-perception-" based control architecture of a humanoid robot. The designed control architecture takes advantage of visual intention allowing the robot some kind of artificial awareness regarding its surrounding environment.

The paper is organized in five sections. The next section briefly introduces the multilevel cognitive concept. Section 3 describes the general structure of a cognitive function and depicts the suggested motion-perception control strategy. Section 4 presents the visual intention bloc. Validation results, obtained from implementation on a real humanoid-like robot, are reported in this section. Finally, the last section concludes the paper.

2. Brief Overview of Multilevel Cognitive Concept

Within the frame of this concept, we consider a process (mainly a complex process) as a multimodel structure where involved components (models), constructed as a result of machine learning (ML), handle two categories of operational levels: reflexive and intentional [21]. This means that ML and related techniques play a central role in this concept and the issued architectures. According to what has been mentioned in the introductory section, two kinds of functions, so-called UCFs and CCFs, build up functional elements ruling the complex task or the complex behavior. Figure 1 illustrates the bloc diagram of the proposed cognitive conception. As it is noticeable from this figure, within the proposed concept,

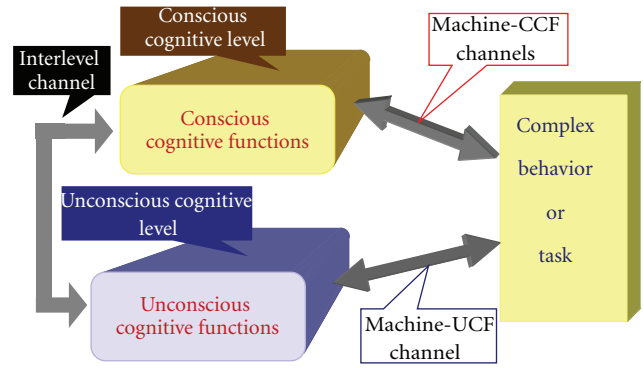

Figure 1: Robot coordinates described by a triplet as $P(x, y, \theta)$.

the overall architecture is obtained by building up cognitive layers (levels) corresponding to different skills fashioning the complex task. It is pertinent to remind that, as well as UCFs, CCFs enclose a number of "elementary functions" (EFs). Within such a scheme, a cognitive layer may fulfil a skill either independently of other layers (typically, the case of unconscious cognitive levels) or using one or several talents developed by other layers (characteristically, the case of conscious cognitive levels) [21].

The first key advantage of conceptualizing the problem within such an incline is to detach the modelling of robots complex artificial behaviours from the type of robot. In other words, models built within such conceptualizing could be used for modelling the same kind of complex behaviours for different kinds of robots. An example of analogy (similarity) with natural cognitive mechanisms could be found in early-ages human walking development. In fact, in its global achievement, the early-ages human abilities development does not depend on the kind of "baby." The second chief benefit of such a concept is that the issued artificial structures are based on "machine learning" paradigms (artificial neural networks, fuzzy logic, reinforcement learning, etc.), taking advantage of "learning" capacity and "generalization" propensity of such approaches. This offers a precious potential to deal with high dimensionality, nonlinearity, and empirical (non-analytical) proprioceptive or exteroceptive information.

3. From Cognitive Function to Motion-Perception Architecture

As it has been mentioned above, a cognitive function (either UCF or CCF) is constructed by a number of EFs. EF is defined as a function (learning-based or conventional) realizing an operational aptitude composing (necessary for) the skill accomplished by the concerned cognitive function. An EF is composed of "elementary components" (ECs). An EC is the lowest level component (module, transfer function, etc.) realizing some elementary aptitude contributing in EF operational aptitude. Two kinds of ECs could be defined (identified): the first corresponding to elementary action that we call "action elementary component" (AEC) and

the second corresponding to elementary decision that we call "decision elementary component" (DEC). An EF may include one or both kinds of the above-defined EC. In the same way, a cognitive function may include one or several ECs. Figure 2 gives the general structure of a cognitive function. However, it is pertinent to notice that there is any restriction to the fact that when it may be necessary, an EC could play the role of an EF. In the same way, when necessary, a cognitive function could include only one EF.

Supposing that a given cognitive function (either conscious or unconscious) includes K ($K \in N$, where N represents the "natural numbers ensemble) elementary functions, considering the kth EF (with $k \in N$ and $k \leq K$) composing this cognitive function, we define the following notations.

Ψ_k is the input of kth EF: $\Psi_k = [\psi_1, \ldots, \psi_j, \ldots, \psi_M]^T$, where ψ_j represents the input component of the jth EC of this EF, $j \leq M$, and M the total number of elementary components composing this EF.

O_k is the output of kth EF.

o_j is the output of the jth EC of the kth EF, with $j \leq M$, and M the total number of elementary components composing the kth EF.

$F_k(\cdot)$ is the skill performed by the kth EF.

$f_j^A(\cdot)$ is the function (transformation, etc.) performed by jth AEC.

$f^D(\cdot)$ is the decision (matching, rule, etc.) performed by DEC.

Within the above-defined notation, the output of kth EF is formalized as shown in (1) with o_j given by (2). In a general case, the output of an EC may also depend on some internal (specific) parameters particular to that EC [21]:

$$O_k = F_k(\Psi_k) = f^D\left(\Psi_k, o_1, \ldots, o_j, \ldots, o_M\right), \qquad o_j = f_j^A\left(\psi_j\right). \tag{1}$$

Based on the aforementioned cognitive concept, the control scheme of a robot could be considered within the frame of "motion-perception-" (MP-) based architecture. Consequently, as well as the robot motions its perception of the environment is obtained combining UCF and CCF. Robot sway is achieved combining unconscious and conscious cognitive motion functions (UCMFs and CCMFs, resp.). In the same way, essentially based on vision, robot perceptual ability is constructed combining unconscious and conscious cognitive visual functions (UCVFs and CCVFs, resp.). Figure 3 shows such an MP-based robot cognitive control scheme. It is pertinent to notice that the proposed control scheme takes advantage of some universality, conceptualizing the build-up of both robot motion and perception abilities independently of the type of robot. It is also relevant to emphasize that the proposed cognitive

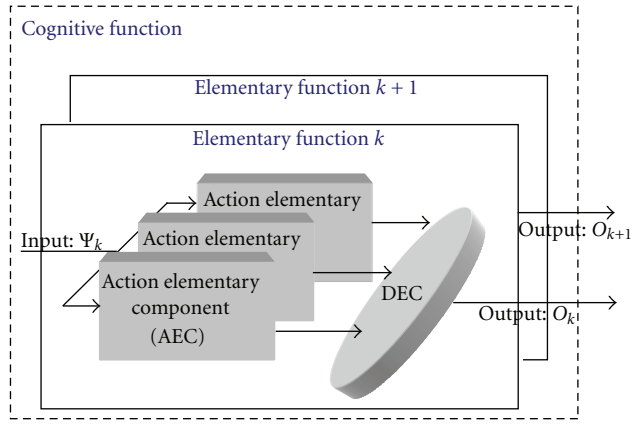

Figure 2: General bloc diagram of a cognitive function.

scheme links the behavior control construction to perception constructing the robot action from and with perceptual data and interaction with the context. This slant of view lays the robot way of doing (e.g., robot knowledge construction) to the human way of learning and knowledge construction: humans or animals learn and construct the knowledge by interacting with the environment. In other words, these natural intelligent beings operate using "awareness" about the surrounding environment in which they live.

If the question of how humans learn, represent, and recognize objects under a wide variety of viewing conditions is still a great challenge to both neurophysiology and cognitive researchers [22], a number of works relating the human early-ages cognitive walking ability construction process highlighting a number of key mechanisms. As shows clinical experiments (as those shown by [23]), one them is the strong linkage between visual and motor mechanisms. This corroborates the pertinence of the suggested cognitive MP-based scheme. Beside this, [24, 25] show that apart of shaping (e.g., recognizing objects and associating shapes with them), we (human) see the world by bringing our attention to visually important objects first. This means that the visual attention mechanism plays also one of the key roles in human infants learning of the encountered objects. Thus, it appears appropriate to draw inspiration from studies on human infants visual learning in constructing robots awareness on the basis of learning by visual revelation.

Making an intelligent system perceive the environment in which it evolves and construct the knowledge by learning unknown objects present in that environment makes a clear need appear relating the ability to select from the overwhelming flow of sensory information only the pertinent ones. This foremost ability is known as "visual saliency," sometimes called in the literature "visual attention," unpredictability, or surprise. It is described as a perceptual quality that makes a part of an image stand out relative to the rest of the image and to capture attention of observer [26]. It may be generalized that it is the saliency (in terms of motion, colors, etc.) that lets the pertinent information "stand out" from the context [27]. We argue that in this context visual saliency may be helpful to enable unsupervised extraction and subsequent learning of

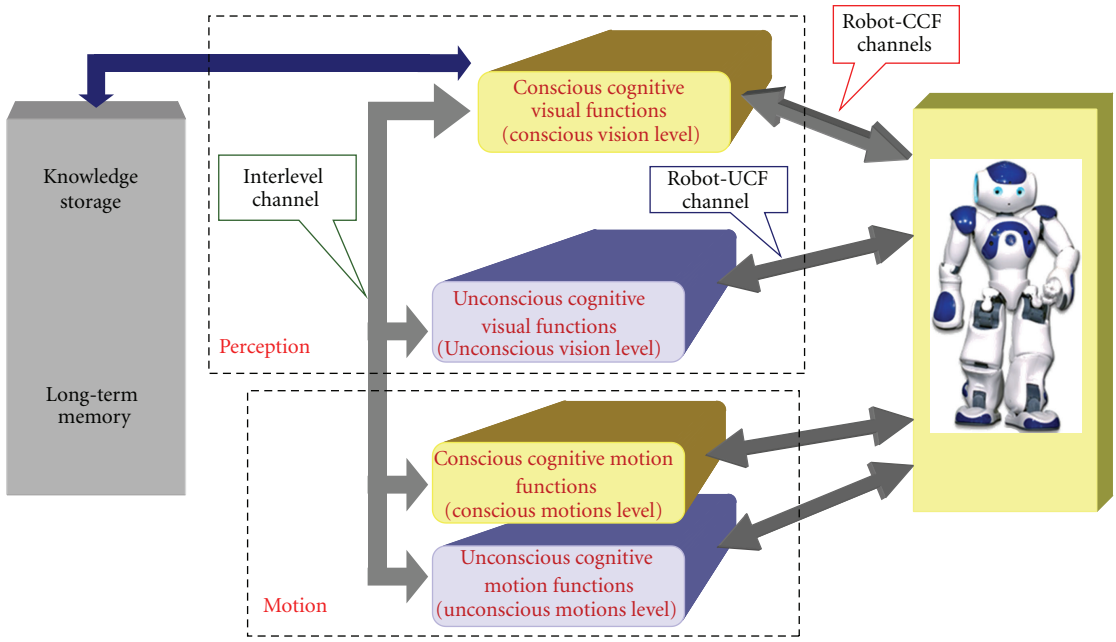

Figure 3: Bloc diagram of motion-perception-based robot cognitive control scheme.

a previously unknown object by a machine, in other words, proffering to the machine (robot) the awareness about its environment.

Referring to the perception bloc of Figure 3, the visual perception is composed of an unconscious visual level including UCVF and one conscious visual level containing CCVF. Unconscious visual level handles reflexive visual tasks, namely, the preprocessing of acquired images, the salient objects detection, and the detected salient objects storage. If the preprocessing could appear as an independent UCVF, it also may be an EF of one of UCVFs composing the unconscious visual level. In this second way of organizing the unconscious visual level, the UCVF including the preprocessing task will deliver the preprocessing results (as those relating image segmentation, different extracted features, etc.) as well to other UCVFs composing the unconscious level as to those CCVFs of conscious level which need the aforementioned results, using the interlevel channel. Conscious visual level conducts intentional visual tasks, namely, the objects learning (including learning detected salient objects), the knowledge construction by carrying out an intentional storage (in unconscious visual level) of new detected salient objects, the detected salient objects recognition in robot surrounding environment (those already known and the visual target (recognized salient object) tracking) allowing the robot self-orientation and motion toward a desired recognized salient object. Consequently, the conscious level communicates (e.g., delivers the outputs of concerned CCVF) with unconscious level (e.g., to the concerned UCVF) as well as with unconscious motion and conscious motion levels (e.g., with the bloc in MP-based robot cognitive control scheme in charge of robot motions).

4. From Salient Objects Detection to Visual Awareness

This section is devoted to description of two principle cognitive visual functions. The first subsection will detail the main UCVF, called "salient vision," which allows robot to self-discover (automatically detect) pertinent objects within the surrounding environment. While, the second subsection will spell out one of the core CCVFs, called "visual intention," which proffers the robot artificial visual intention ability and allows it to construct the knowledge about the surrounding environment proffering the robot the awareness regarding its surrounding environment.

Before describing the above-mentioned functions, it is pertinent to note that a recurrent operation in extracting visually salient objects (from images) is image segmentation. Generally speaking, one can use any available image segmentation technique. However, the quality of segmentation may be weighty for an accurate extraction of salient objects in images. In fact, most of the usual segmentation techniques (used beside standard image salient object extraction techniques) using manual or automatic thresholding remain limited because they do not respect the original image features. That is why we made use of the algorithm proposed recently by [28]. It is based on K-means clustering of color space with an adaptive selection of K and a spatial filter removing meaningless segments. The used algorithm is very fast (tens of milliseconds for a 320×240 pixels image on a standard PC) and it claims to have results close to human perception. The execution speed is a major condition in effective implementation in robotics applications reinforcing our choice for this already available algorithm, which keeps upright between execution speed and achieved segmentation quality.

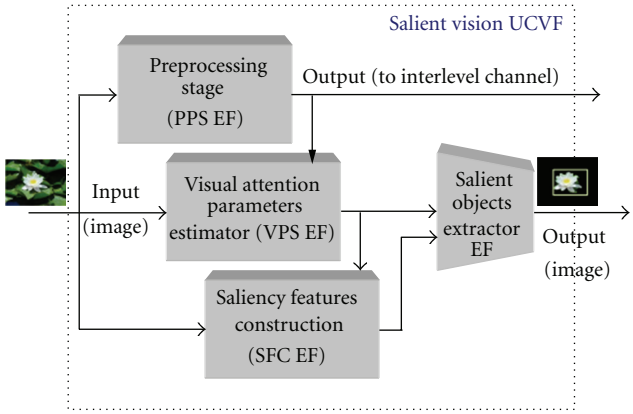

Figure 4: Bloc-diagram of Salient Vision UCVF, handling the automated detection of salient objects.

4.1. Salient Vision UCVF and Salient Objects Detection. The bloc diagram detailing the structure of "salient vision" UCVF is given in Figure 4. As it is visible from this figure, the "salient vision" UCVF includes also the preprocessing stage (defined as one of its constituting EFs), meaning that this UCVC handles the image segmentation and common image features' extraction tasks, delivering the issued results to other UCVFs as well as to conscious visual level. Beside this EF, it includes three other EF: the "visual Attention parameters estimator" (VAPE) EF, the "salient features construction" (SFC) EF, and the "salient objects extraction" (SOE) EF. This last EF plays the role of a decision-like elementary component, implemented as an independent EF.

4.1.1. Visual Attention Parameters Estimation Elementary Function. The visual attention parameter estimation (VAPE) elementary function determines what could be assimilated to some kind of "visual attention degree." Computed on the basis of preprocessing bloc issued results and controlling local salient features, visual attention parameter p constructs a top-down control of the attention and of the sensitivity of the feature in scale space. High value of p (resulting in a large sliding window size) with respect to the image size will make the local saliency feature more sensitive to large objects. In the same way, low values of p allow focusing the visual attention on smaller objects and details. The value of visual attention parameter p can be hard-set to a fixed value based on a heuristic according to [29]. However, as different images usually present salient objects in different scales, this way of doing will limit the performance of the system. Thus, a new automated cognitive estimation of the parameter p has been designed. The estimation is based, on the one hand, on calculation (inspired from the work presented in [30]) of the histogram of segment sizes from the input image, and on the other hand, on using of an artificial neural network (ANN). The ANN receives (as input) the feature vector issued from the above-mentioned histogram and provides the sliding window value. The weights of the neural network are adapted in training stage using a genetic algorithm.

To obtain the aforementioned histogram, the input image is segmented into n segments $(S_1, S_2, \ldots, S_n)$. For each one of the found segments S_i (where $S_i \in \{S_1, S_2, \ldots, S_n\}$), its size $|S_i|$ (measured in number of pixels) is divided by the overall image size $|I|$. An absolute histogram H_{SA} of segment sizes is constructed according to (2), avoiding leading to a too sparse histogram. This ensures that the first histogram bin contains the number of segments with area larger than 1/10 of the image size, the second contains segments from 1/10 to 1/100 of the image size, and so forth. For practical reasons we use a 4-bin histogram. Then, this absolute histogram leads to a relative histogram H_{SR} computed according to relation (3):

$$H_{SA}(i) = \sum_{j=1}^{n} \begin{cases} 1 & \text{if } 10^{i-1} \leq \dfrac{|S_i|}{|I|} \leq 10^{i}, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

$$H_{SR}(i) = \frac{H_{SA}(i)}{\sum_j H_{SA}(j)}. \tag{3}$$

The core of the proposed visual attention parameter estimator is a fully connected three-layer feed-forward MLP-like ANN, with a sigmoidal activation function, including 4 input nodes, 3 hidden neurons, and 1 output neuron. The four input nodes are connected each to its respective bin from the H_{SR} histogram. The value of the output node, belonging to the continuous interval $[0, 1]$, could be interpreted as the ratio of the estimated sliding window size p and the long side size of the image. The ANN is trained making use of a genetic algorithm described in [31]. Each organism in the population consists of a genome representing an array of floating point numbers whose length corresponds with the number of weights in MLP. To calculate the fitness of each organism, the MLP weights are set according to its current genome. Once visual attention parameter p is available (according to the MLP output) saliency is computed over the image and salient objects are extracted. The result is compared with ground truth and the precision, the recall and the F-ratio (representing the overall quality of the extraction) are calculated (according to [32] and using the measures proposed in the same work to evaluate quantitatively the salient object extraction). The F-ratio is then used as the measure of fitness. In each generation, the elitism rule is used to explicitly preserve the best solution found so far. Organisms are mutated with 5% of probability. As learning data set, we use 10% of the MSRA-B data set (described in [32]). The remaining 90% of the above-indicated data set has been used for validation.

4.1.2. Salient Features Construction Elementary Function. The salient features construction (SFC) elementary function performs two kinds of features (both used for salient objects detection). The first kind is global saliency features and the second local saliency features. Global saliency features capture global properties of image in terms of distribution of colors. The global saliency is obtained combining "intensity saliency" and the "chromatic saliency." Intensity saliency $M_l(x)$, given by relation (4), is defined as Euclidean distance of intensity I to the mean of the entire image. Index l stands

Figure 5: Examples of global and local saliency features: original image (a), global saliency map (b), local saliency map (c), and final saliency map.

for intensity channel of the image, and $I_{\mu l}$ is the average intensity of the channel. In the same way, chromatic saliency, given by relation (6), is defined as Euclidean distance of azimuth and zenith components intensities (e.g., azimuth ϕ and zenith θ, resp.) to their means ($I_{\mu\phi}$ and $I_{\mu\theta}$ resp.) in the entire image. Term (x) denotes coordinates of a given pixel on the image:

$$M_l(x) = \left\| I_{\mu l} - I_l(x) \right\|,$$
$$M_{\phi\theta}(x) = \sqrt{\left(I_{\mu\phi} - I_\phi(x)\right)^2 + \left(I_{\mu\theta} - I_\theta(x)\right)^2}. \tag{4}$$

The global saliency map $M(x)$, given by relation (5) is a hybrid result of combination of maps resulted from (1) according to logistic sigmoid blending function. Blending of the two saliency maps together is driven by a function of color saturation C of each pixel. It is calculated from RGB color model for each pixel as pseudonorm, given by $C = \mathrm{Max}[\mathrm{R}, \mathrm{G}, \mathrm{B}] - \mathrm{Min}[\mathrm{R}, \mathrm{G}, \mathrm{B}]$. When C is low, importance is given to intensity saliency. When C is high, chromatic saliency is emphasized:

$$M(x) = \frac{1}{1 - e^{-C}} M_{\phi\theta}(x) + \left(1 - \frac{1}{1 + e^{-C}}\right) M_l(x). \tag{5}$$

The global saliency (and related features) captures the visual saliency with respect to the colors. However, in real cases, the object visual saliency may also consist in its particular shape or texture, distinct to its surroundings, either beside or rather than simply in its color. To capture this aspect of visual saliency, a local feature over the image is determined. Inspired from a similar kind of feature introduced in [32], the local saliency has been defined as a centre-surround difference of histograms. The idea relating the local saliency 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 within the sliding window.

To formalize this idea leading local saliency features, let us have a sliding window P of size p, centered over pixel (x). Define a (centre) 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 centre-surround feature $d(x)$ is then given as (6) over all histogram bins (i):

$$d(x) = \sum_i \frac{|H_C(i) - H_S(i)|}{p^2}. \tag{6}$$

Resulting from computation of the $d(x)$ throughout all the l, ϕ, and θ channels, the centre-surround saliency $D(x)$ on a given position (x) is defined according to (7). Similarly to (5), a logistic sigmoid blending function has been used to combine chromaticity and intensity in order to improve the performance of this feature on images with mixed achromatic and chromatic content. However, here the color saturation C refers to average saturation of the content of the sliding window P:

$$D(x) = \frac{1}{1 - e^{-C}} d_l(x) + \left(1 - \frac{1}{1 + e^{-C}}\right) \mathrm{Max}\left(d_\phi(x), d_\theta(x)\right). \tag{7}$$

4.1.3. Salient Objects Extraction Elementary Function. Salient objects extraction (SOE) elementary function acts as the last step of saliency map calculation and salient objects detection. The extracted global and local salient features (e.g., $M(x)$ and $D(x)$, resp.) are combined using (8), resulting in final saliency map $M_{\mathrm{final}}(x)$, which is then smoothed by Gaussian filter. The upper part of the condition in (8) describes a particular case, where a part of image consists of a color that is not considered salient (i.e., pixels with low $M(x)$ measure) but which is distinct from the surroundings by virtue of its shape.

(a) (b) (c)

Figure 6: Examples of salient object detection: input image (a), detected salient objects (b), and ground truth salient objects (c).

(a) (b) (c)

Figure 7: Effect of the visual attention parameter p: input image (a), detected salient objects with high values of p (b) and small values of p (c).

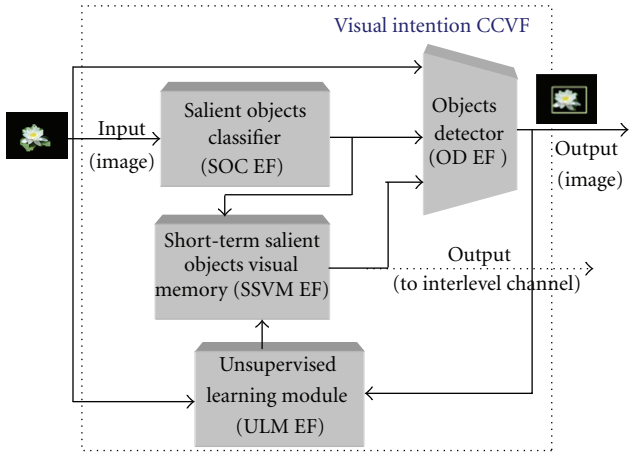

Figure 8: Bloc diagram of visional intention CCVF.

The final saliency map samples are shown on the column d of Figure 5:

$$M_{\text{final}}(x) = \begin{cases} D(x) & \text{if } M(x) < D(x), \\ \sqrt{M(x)D(x)} & \text{otherwise.} \end{cases} \quad (8)$$

Accordingly to segmentation and detection algorithms described in [30, 33], the segmentation splits an image into a set of chromatically coherent regions. Objects present on the scene are composed of one or multiple such segments. For visually salient objects, 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. Conformably to [33], 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 (where $S_i \in \{S_1, S_2, \ldots, S_n\}$), its average saliency $\overline{S_i}$ and variance (of saliency values) $\text{Var}(S_i)$ are computed over the final saliency map $M_{\text{final}}(x)$. All the pixel values $p(x, y) \in S_i p(z, y)$ of the segment are then set following (9), where $\tau_{\overline{S_i}}$ and τ_{Var} are thresholds for average saliency 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 binary value "1". To remove noise, a membership condition is imposed that any $C_i \in C$ has its area larger than a given threshold. Finally, the binary map is projected on the original image leading to a result that is part (areas) of the original image containing its salient objects. References [33, 34] give different values for the aforementioned parameters and thresholds:

$$p(x, y) = \begin{cases} 1 & \text{if } \overline{S_i} > \tau_{\overline{S_i}},\ \text{Var}(S_i) > \tau_{\text{Var}}, \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Figure 5 shows examples of global and local saliency features extracted from two images. In the first image the global salient feature (upper image of column b) is enough to track salient objects, while for the second, where the salient object (leopard) is partially available, chromatic saliency is not enough to extract the object. Figures 6 and 7 show examples

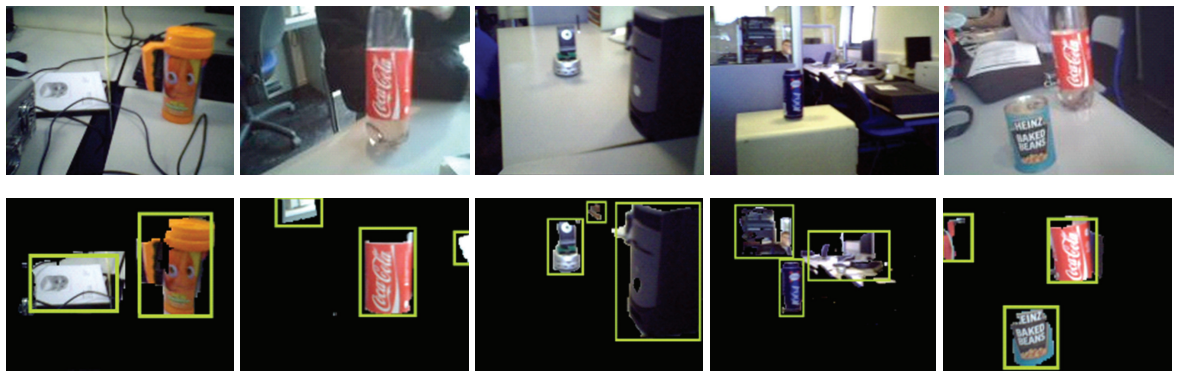

Figure 9: NAO robot camera issued images (upper images) and corresponding salient objects found and segmented by NAO (lower images).

Figure 10: Results relative to a set of objects detection by robot.

of salient object detection as well as effect of the visual attention parameter p on extracted salient regions, respectively.

4.2. Visual Intention CCVF. As it has previously been stated, composed of conscious cognitive visual functions (CCVFs), the conscious visual level conducts intentional visual tasks. One of the core functions of this level is "visual intention" CCVF, proffering the robot some kind of "artificial visual intention ability" and allowing the machine to construct its first knowledge about the surrounding environment. Figure 8 gives the bloc diagram of visional intention CCVF. As it could be seen from this figure, this CCVF is composed of four elementary functions: "short-term salient objects visual memory" (SSVM) EF, "unsupervised learning module" (ULM) EF, "salient objects classifier" (SOC) EF, and "object detector" (OD) EF.

The main task of short-term salient objects visual memory (SSVM) EF is to provide already known objects and store currently recognized or detected salient objects. It could also be seen as the first knowledge construction of surrounding environment because it contains the clusters of salient objects resulting from unsupervised learning. Its content (e.g., stored salient objects or groups of salient objects) could supply the main knowledge base (a long-term memory). That is why its output is also connected to interlevel channel.

The role of unsupervised learning (performed by ULM EF) is to cluster the detected (new) salient objects. The learning process is carried out on line. When an agent (e.g., robot) takes images while it encounters a new object, if the objects are recognized to be salient (e.g., extracted) they are grouped incrementally while new images are acquired. The action flow of the learning process is given below. In the first time, the algorithm classifies each found fragment, and, in a second time, the learning process is updated (online learning)

```
acquire image
extract fragments by salient object detector
for each fragment F
  if(F is classified into one group)
    populate the group by F
  if(F is classified into multiple groups)
    populate by F the closest group by Euclidian distance of features
    if(F is not classified to any group)
        create a new group and place F inside
select the most populated group G
use fragments from G as learning samples for object detection algorithm
```

The salient objects classifier is a combination of four weak classifiers $\{w_1, w_2, w_3, w_4\}$, each classifying a fragment as belonging or not belonging to a certain class. F denotes the currently processed fragment, and G denotes an instance of the group in question. The first classifier w_1, defined by (10), separates fragments with too different areas. In experiments $t_{\text{area}} = 10$. The w_2, defined by (11), separates fragments whose aspects are too different to belong to the same object.

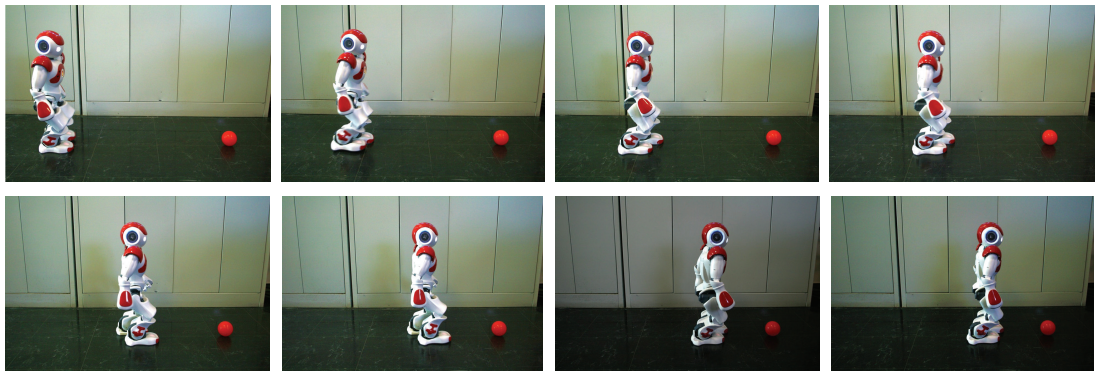

Figure 11: Results relative to an intentional object tracking: the robot tracks a red ball moving toward it.

In experiments, t_{aspect} has been set to 0.3. The classifier w_3, defined by (12), separates fragments with clearly different chromaticity. It works over 2D normalized histograms of ϕ and θ component denoted by $G_{\phi\theta}$ and $F_{\phi\theta}$, respectively, with L bins, calculating their intersection. We use $L = 32$ to avoid too sparse histogram and $t_{\phi\theta}$ equal to 0.35. Finally, w_4 (defined by (13)) separates fragments whose texture is too different. We use the measure of texture uniformity calculated over the l channel of fragment. $p(z_i)$, where $i \in \{0, 1, 2, \ldots, L-1\}$, is a normalized histogram of l channel of the fragment, and L is the number of histogram bins. In experiments, 32 histogram bins have been used to avoid too sparse histogram and value $t_{\text{uniformity}}$ of 0.02. A fragment belongs to a class if $\prod_{i=1}^{n} w_i = 1$:

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

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

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

$$w_4 = \begin{cases} 1 & \text{if } c_{w4} < t_{\text{uniformity}}, \\ 0 & \text{otherwise,} \end{cases}$$
$$c_{w4} = \left\| \sum_{j=0}^{L-1} p_G^2(z_j) - \sum_{k=0}^{L-1} p_F^2(z_k) \right\|. \tag{13}$$

4.3. Implementation on Real Robot and Experimental Validation. The above-described concept has been implemented on NAO robot, which includes vision devices and a number of onboard preimplemented motion skills. It also includes a number of basic standard functions that have not been used. For experimental verification, the robot has been introduced in a real environment with different common objects (representing different surface, shapes, and properties). Several objects were exposed in robots field of view, presented in a number of contexts different from those used in the learning phase. The number of images acquired for each object varied between 100 and 600 for learning sequences and between 50 and 300 for testing sequences, with multiple objects occurring on the same scene. During the learning process, the success rate of 96% has been achieved concerning pertinent learned objects (e.g., those identified by the robot as salient and then learned), that is, only 4% of image fragments were associated with wrong groups. During the testing process, objects were correctly extracted reaching 82% success rate.

To demonstrate real-time abilities of the system, the NAO robot was required to find some learned objects in its environment and then to track them. It is pertinent to emphasize that those objects have been learned in different environment. A sample of results of those experiments is shown in Figures 9 to 12. Figures 9 and 10 show results relating robot ability to detect and extract salient objects from its surrounding environment. It is pertinent to notice the multiple salient objects detection ability of the implemented strategy representing different shapes and various natures. Figure 11 shows the expected robot ability to detect and to follow a simple object in real environment, validating the correct operation of unconscious and intentional cognitive levels transitions in accomplishing the required task. Finally, Figure 12 shows the robot ability to detect, isolate, and follow a previously detected and learned salient object in a complex surrounding environment. The video of this experiment could be seen using the link indicated in the legend of this figure. It is pertinent to emphasize the fact that the object (a "book" in the experiment shown by the Figure 12) has been detected and learned in different conditions (as one could see this from the above-indicated video). Thus, this experiment shows the emergence of a kind of robot "artificial awareness" about the surrounding environment validating the presented cognitive multilevel concept and issued "perception-motion" architecture.

Figure 12: Tracking a previously learned moving object (upper images: video http://www.youtube.com/watch?v=xxz3wm3L1pE). The upper right corner of each image shows robot camera picture.

5. Conclusion

By supplanting the modeling of robots complex behavior from the "control theory" backdrop to the "cognitive machine learning" backcloth, the proposed machine-learning-based multilevel cognitive motion-perception concept attempts to offer a unified model of robot autonomous evolution, slotting in two kinds of cognitive levels: "unconscious" and "conscious" cognitive levels, answerable of its reflexive and intentional visual and motor skills, respectively.

The first key advantage of conceptualizing the problem within such incline is to detach the build-up of robot perception and motion from the type of machine (robot). The second chief benefit of the concept is that the issued structure is "machine-learning-" based foundation taking advantage from "learning" capacity and "generalization" propensity of such models.

The "visual intention" built-in CCVF proffers the robot artificial visual intention ability and allows it to construct the knowledge about the surrounding environment. This intentional cognitive function holds out the robot awareness regarding its surrounding environment. The extracted knowledge is first stored in (and recalled from) short-term memory. It could then be stored in a long-term memory proffering the robot some kind of learning issued knowledge about previously (already) explored environments or already known objects in a new environment. Beside this appealing ability, the unconscious visual level realizing the salient objects detection plays a key role in the so-called "artificial awareness" emergence. In fact, the ability of automatic detection of pertinent items in surrounding environment proffers the robot some kind of "unconscious awareness" about potentially significant objects in the surrounding environment. The importance of this key skill appears not only in emergence of "intentional awareness" but also in construction of new knowledge (versus the already learned items) upgrading the robot (or machine's) awareness regarding its surrounding environment.

References

[1] E. R. Westervelt, G. Buche, and J. W. Grizzle, "Experimental validation of a framework for the design of controllers that induce stable walking in planar bipeds," *International Journal of Robotics Research*, vol. 23, no. 6, pp. 559–582, 2004.

[2] J. H. Park and O. Kwon, "Reflex control of biped robot locomotion on a slippery surface," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '01)*, pp. 4134–4139, May 2001.

[3] J. Chestnutt and J. J. Kuffner, "A tiered planning strategy for biped navigation," in *Proceedings of the 4th IEEE-RAS International Conference on Humanoid Robots (Humanoids '04)*, vol. 1, pp. 422–436, November 2004.

[4] Q. Huang, K. Yokoi, S. Kajita et al., "Planning walking patterns for a biped robot," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 280–289, 2001.

[5] K. Sabe, M. Fukuchi, J. S. Gutmann, T. Ohashi, K. Kawamoto, and T. Yoshigahara, "Obstacle avoidance and path planning for humanoid robots using stereo vision," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '4)*, pp. 592–597, May 2004.

[6] R. Holmes, *Acts of War: The Behavior of Men in Battle*, The Free Press, New York, NY, USA, 1st American edition, 1985.

[7] M. Tambe, W. L. Johnson, R. M. Jones et al., "Intelligent agents for interactive simulation environments," *AI Magazine*, vol. 16, no. 1, pp. 15–40, 1995.

[8] P. Langley, "An abstract computational model of learning selective sensing skills," in *Proceedings of the 18th Conference of the Cognitive Science Society*, pp. 385–390, 1996.

[9] C. Bauckhage, C. Thurau, and G. Sagerer, "Learning human-like opponent behavior for interactive computer games," *Lecture Notes in Computer Science*, vol. 2781, pp. 148–155, 2003.

[10] V. Potkonjak, D. Kostic, S. Tzafestas, M. Popovic, M. Lazarevic, and G. Djordjevic, "Human-like behavior of robot arms: general considerations and the handwriting task—part II: the robot arm in handwriting," *Robotics and Computer-Integrated Manufacturing*, vol. 17, no. 4, pp. 317–327, 2001.

[11] J. Edlund, J. Gustafson, M. Heldner, and A. Hjalmarsson, "Towards human-like spoken dialogue systems," *Speech Communication*, vol. 50, no. 8-9, pp. 630–645, 2008.

[12] A. Lubin, N. Poirel, S. Rossi, A. Pineau, and O. Houdé, "Math in actions: actor mode reveals the true arithmetic abilities of french-speaking 2-year-olds in a magic task," *Journal of Experimental Child Psychology*, vol. 103, no. 3, pp. 376–385, 2009.

[13] F. A. Campbell, E. P. Pungello, S. Miller-Johnson, M. Burchinal, and C. T. Ramey, "The development of cognitive and academic abilities: growth curves from an early childhood educational experiment," *Developmental Psychology*, vol. 37, no. 2, pp. 231–242, 2001.

[14] G. Leroux, M. Joliot, S. Dubal, B. Mazoyer, N. Tzourio-Mazoyer, and O. Houdé, "Cognitive inhibition of number/length interference in a Piaget-like task in young adults: evidence from ERPs and fMRI," *Human Brain Mapping*, vol. 27, no. 6, pp. 498–509, 2006.

[15] A. Lubin, N. Poirel, S. Rossi, C. Lanoë, A. Pineau, and O. Houdé, "Pedagogical effect of action on arithmetic performances in Wynn-like tasks solved by 2-year-olds," *Experimental Psychology*, vol. 57, no. 6, pp. 405–411, 2010.

[16] O. C. S. Cassell, M. Hubble, M. A. P. Milling, and W. A. Dickson, "Baby walkers—still a major cause of infant burns," *Burns*, vol. 23, no. 5, pp. 451–453, 1997.

[17] M. Crouchman, "The effects of babywalkers on early locomotor development," *Developmental Medicine and Child Neurology*, vol. 28, no. 6, pp. 757–761, 1986.

[18] A. Siegel and R. Burton, "Effects of babywalkers on early locomotor development in human infants," *Developmental & Behavioral Pediatrics*, vol. 20, pp. 355–361, 1999.

[19] I. B. Kauffman and M. Ridenour, "Influence of an infant walker on onset and quality of walking pattern of locomotion: an electromyographic investigation," *Perceptual and Motor Skills*, vol. 45, no. 3, pp. 1323–1329, 1977.

[20] J. A. Andersen, *An Introduction to Neural Network*, MIT Press, Cambridge, Mass, USA, 1995.

[21] K. Madani and C. Sabourin, "Multi-level cognitive machine-learning based concept for human-like "artificial" walking: application to autonomous stroll of humanoid robots," *Neurocomputing*, vol. 74, no. 8, pp. 1213–1228, 2011.

[22] H. Bülthoff, C. Wallraven, and M. Giese, "Perceptual robotic," in *Handbook of Robotics*, B. Siciliano and O. Khatib, Eds., Springer, 2007.

[23] http://www.universcience-vod.fr/media/577/la-marche-des-bebes.html.

[24] P. Zukow-Goldring and M. A. Arbib, "Affordances, effectivities, and assisted imitation: caregivers and the directing of attention," *Neurocomputing*, vol. 70, no. 13–15, pp. 2181–2193, 2007.

[25] R. J. Brand, D. A. Baldwin, and L. A. Ashburn, "Evidence for "motionese": modifications in mothers' infant-directed action," *Developmental Science*, vol. 5, no. 1, pp. 72–83, 2002.

[26] R. Achanta, S. Hemami, E. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR '09)*, 2009.

[27] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Reviews Neuroscience*, vol. 5, no. 6, pp. 495–501, 2004.

[28] T. W. Chen, Y. L. Chen, and S. Y. Chien, "Fast image segmentation based on K-means clustering with histograms in HSV color space," in *Proceedings of the IEEE 10th Workshop on Multimedia Signal Processing (MMSP '08)*, pp. 322–325, October 2008.

[29] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, vol. 2, pp. 1–8, June 2007.

[30] R. Moreno, M. Graña, D. M. Ramik, and K. Madani, "Image segmentation by spherical coordinates," in *Proceedings of the 11th International Conference on Pattern Recognition and Information Processing (PRIP '11)*, pp. 112–115, 2011.

[31] J. H. Holland, *Adaptation in Natural anti Artificial Systems: An introductory Analysis with Applications to Biology, Control and Artificial Intelligence*, MIT Press, 1992.

[32] T. Liu, Z. Yuan, J. Sun et al., "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[33] D. M. Ramík, C. Sabourin, and K. Madani, "Hybrid salient object extraction approach with automatic estimation of visual attention scale," in *Proceedings of the 7th International Conference on Signal Image Technology & Internet-Based Systems (IEEE—SITIS '11)*, pp. 438–445, 2011.

[34] D. M. Ramík, C. Sabourin, and K. Madani, "A cognitive approach for robots' vision using unsupervised learning and visual saliency," in *Advances in Computational Intelligence*, vol. 6691 of *LNCS*, pp. 65–72, Springer, 2011.