Bioinspired Early Visual Processing: the Attention Condensation Mechanism

Thomas Müller, Alois Knoll
Robotics and Embedded Systems,
Technische Universtät München,
Boltzmannstr. 3, 85748 Garching, Germany
{muelleth,knoll}@cs.tum.edu

Abstract

In this paper we propose an biologically inspired attention-based vision system for the JAST interactive dialog robot. The robotvision system incorporates three submodules: object recognition, gesture recognition and self recognition.

Herein two assumptions form the theoretical foundation: first and generally, attention is effected by bottom-up attractors. These may arise from high intensity / hue gardients or scene dynamics. Second, the focus of attention can be directed by higher level processes, whether volitionally or not, in an inhibitory or reinforcing way. The system proposed in this paper utilizes these assumptions to organize its computational efforts in an *Attention Condensation Layer* accordingly.

Due to its efficient data management architecture, the system is capable of continuously publishing results to the robot's cognitive layer and thus operating in realtime. Furthermore, the modular structure and the asynchronous communication paradigm allows for efficient integration of additional modules, be it visual or from any other sensor.

The main contribution of this work is the application of neuroscientific findings on human early visual processing to a real-world robotic setup. Here, our experimental results show tremendous speed-ups compared to naive implementations, reaching the peak in a combination of the top-down and bottom-up principles.

1 Introduction

Many traditional computer vision systems, either for surveillance tasks with fixed cameras, robot vision on mobile platforms or any other kind of visual processing suffer from an enormous computational complexity. The main reason for this is, that these traditional systems utilize *complete analysis* approaches: after aquiring input data from its camera(s), the system has to process the huge amount of data and analyze all of it regardless of the significance to the current task or environment.

Now, the basic, biologically inspired idea is to apply an attention-based information filter in order to reduce the amount of input data and only perform further analysis on the rest. This residual is what we call the regions of interest (ROIs).

1.1 Related Work

There have been many approaches to computation of salient features in a static image, e.g. [Reinagel and Zador, 1999] show that high contrast regions seem to attract attention or [Kadir and Brady, 2000] report that salient regions can be computed using multiscale images. [Gilles, 1998] on the other hand argues that local complexity can be a measure of saliency. Also, a learning approach for visual saliency models has been proposed recently in [Kienzle *et al.*, 2006]. Following these ideas we claim, that fundamental attention attractors originating from sensory input can be either static salient features in a single frame or dynamics in the input data sequence (considering temporal properties). Details on this topic are addressed in Section 2.

Furthermore, inspired by the idea of [O'Regan and Noë, 2001], which they claim to be biologically plausible, we extend the saliency attention approach with the idea, that vision is a process of active and sometimes even volitional exploration of the subject's environment. Also considering the theory of *inhibition of return*, which was shown to be plausible in human visual psychophysics e.g. by [Posner and Cohen, 1984], we derive our second before mentioned assumption and implement top-down cognitive feedback in the proposed system. Moreover, we integrate the capability not only for inhibition, but also for directed attention guidance. This reinforcement is triggered by cognitive processes reasoning about relevant additional information to gain from a specific region (see Section 3).

Although we do not claim to implement the entire framework of O'Regan's and Noë's theory (e.g. the *inattentional* or *change blindness*), we in deed show that a system utilizing the basic ideas performs considerably better than without.

We are aware of vision systems providing similar capabilities to the one proposed here. E.g. [Itti and Koch, 2001; Itti *et al.*, 1998] implement a visual attention system utilizing multiscale images to compute a saliency map. In their system a neural network selects the attended locations for detailed analysis. [Walther *et al.*, 2002] use a static architecture to perform bottom-up attention based selection and attentional modulation to speed up the recognition process of their connectionist HMAX system.

Not contradicting, but complementing this work, we do not want to focus solely on building a biologically plausible visual systems, but our primary target is to apply the underlying ideas of such frameworks to a real-world robotic setup. We therefore avoid neural, connectionist or machine learning techniques, giving preference to a straight forward implementation of discrete algorithms. These fast and efficient algorithms allow for realtime performance and high accuracy for manipulation tasks on standard hardware.

1.2 The JAST Robot Setup

The vision system presented in this paper is part of a human-robot dialog system, which operates as the main demonstrator platform for the JAST "**J**oint-**A**ction **S**cience and **T**echnology" project.

The overall goal of the JAST project is to investigate the cognitive and communicative aspects of jointly-acting agents, both human and artificial. The human-robot dialog system being built as part of the project [Foster *et al.*, 2007; Rickert *et al.*, 2007] is designed as a platform to integrate the projects empirical findings on cognition and dialog with research on autonomous robots, by supporting symmetrical, multimodal human-robot collaboration on a joint construction task[1].

The robot (Figure 1) consists of a pair of mechanical arms with grippers and an animatronic talking head. The input channels consist of speech recognition, object recognition, gesture recognition, and robot sensors; the outputs include synthesized speech, emotional expressions, head motions, and robot actions. The user and the robot work together to assemble a wooden construction toy on a common work area, coordinating their actions through speech, gestures, and facial motions.

In order to restrict the variety of visual input, vision processing in the JAST system is performed on the output of a single camera which is installed directly above the table looking downward to take images of the scene and the user entering the scene. The camera provides an image stream of 15 frames per second at a resolution of 1024×768 pixels. The output of the vision process is published to the multimodal fusion component [Giuliani and Knoll, 2007], where it is used for disambiguating spoken input from the user. Moreover, combined hypotheses representing the users requests are produced and reasoning on the properties of the observed world parameters is performed.

[1] http://www.youtube.com/watch?v=yXZXnAQ15LI

Figure 1: The JAST human-robot dialog system.

1.3 The Vision Architecture

The vision system presented here (Figure 2) applies an asynchronous communication mechanism (ACM). Therefore we can implement non-blocking behaviour and still guarantee the required frequency for result publishing, as publishing incomplete analysis results is tolerated. Derived from common standards [Message Passing Interface Forum, 1995], intermediate vision data is managed in limited-size priority-queues.

According to e.g. [Sundell and Tsigas, 2003] non-blocking algorithms can be distinguished into being *lock-free* and *wait-free*. Lock-free implementations guarantee at least one process to continue at any time. Wait-free implementations on the other hand avoid starvation as they guarantee completion of a task in a limited number of steps [Herlihy, 1991]. Generally, one can state it essential for systems utilizing an ACM to stay responsive, not to guarantee data transmission.

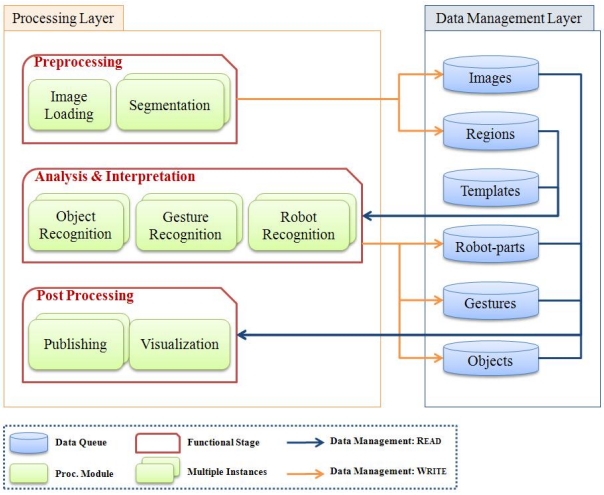

Figure 2: Architectural overview of the vision system.

Concerning parallelization techniques [Culler *et al.*, 1999], the JAST vision system applies a wait-free combination of *data-domain* and *function-domain* parallelization. As previously shown in [Müller *et al.*, 2008], this combined approach performs very well in practice because anchor points for distributed computation can be designed to be independent concerning memory and workflow.

1.4 Analysis and Interpretation

The next paragraphs are dedicated to a brief description of the algorithmics applied within the main components of the vision system, i.e. the *Analysis and Interpretation* stage. The main topic of this paper, corresponding to the *Preprocessing* stage in Figure 2 is then described in Section 2 and Section 3 in great detail.

Object recognition within the scene is fairly straight forward once the regions of interest are identified. We apply the *OpenCV*-implementation of a template-matching algorithm [ope,] on twenty different rotations of each template we consider to be relevant. The templates are generated from previously taken samples (a future version of the system will extract them online).

Since the OpenCV library offers a choice of different similarity measures, we could e.g. utilize cross-correlation (see below) or any other provided patch-comparison method - for example least-square-errors (LSE) or correlation-coefficients.

$$CC_{T,ROI} = \max_{t_{\Delta x,\Delta y,\theta}} \left[\sum_{p \in T} (T(p) \cdot ROI(t(p))) \right] \quad (1)$$

In Equation 1 the transform $t_{\Delta x,\Delta y,\theta}(p)$ denotes the projection of p into the region ROI according to the translation $\Delta x, \Delta y$ (done automatically in OpenCV) and a discrete rotation value $\theta \in [0; 2\pi]$. As we analyze twenty different rotations, we have to call OpenCV's matching function twenty times per template and then take the maximum of the normalized similarities as our result for the template T in region ROI. After comparing the results of each suitable template we select the best ones and finally publish them to the higher level reasoning modules (see Figure 3 for typical match results on two ROIs).

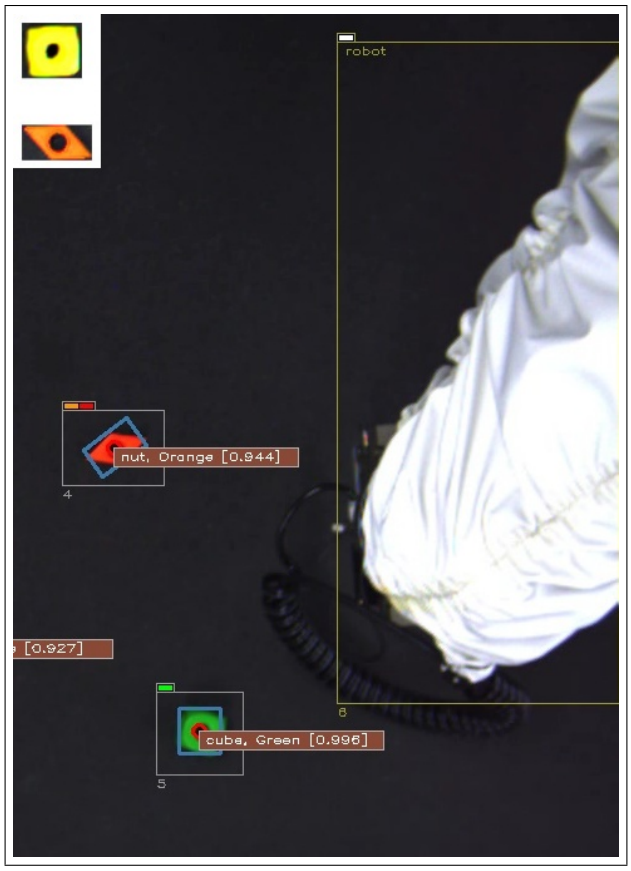

Figure 3: Object recognition and self recognition result for a typical input image (detail). The two templates of highest probability in the template matching process are shown in the left upper corner. The location of the robot is detected utilizing a cognitive feedback algorithm.

Gesture recognition is implemented as a two step approach: first, specific invariants have to be extracted from a region, and second, the gesture has to be classified.

In order to be able to perform a classification, we need to train the system on possible classes of gestures and their specific set of invariant-values in advance. For this task 500 – 600 gesture templates are aquired and tagged with a class label. The training invariants are stored within a vector for each class and used for comparison

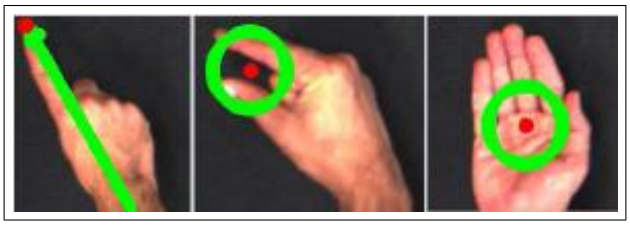

Figure 4: Typical results for the gestures recognized by the JAST vision system. Currently the vision system only interprets the above shown gestures classes.

in the runtime classification.

To classify an extracted set of invariants, we find the K nearest neighbors which are calculated based on the weighted distance of each training vector to the input invariant. The distance from the extracted invariant-values to a training vector can then be computed in Euclidean space.

Next, we choose the K vectors from the training pool which have the shortest normalized distance to the given invariants. A naïve Bayes probability for the class of the given invariants can then be computed for each class of gesture we are able to recognize. Finally the class assigned with the highest probability is published as the result. Figure 4 shows typical results of the recognition process. A detailed description of the gesture recognition can be reviewed from [Ziaie *et al.*, 2008a; 2008b].

Self recognition or robot recognition is acomplished by a cognitive feedback algorithm. From robot sensors one can retrieve the current joint parameters of each joint of the robot arms. This information is used to adjust a 3D model of the robot accordingly. Based on this information, in combination with link properties and the position of the torso, that are known from a priori, the system can compute the 3D cartesian position of each joint applying forward kinematics.

In order get results for self recognition quickly, we simplify the model to a skeleton and add virtual spheres and cylinders of appropriate size around the joints, resp. links. Next, a projection of the calculated model from the 3D space into 2D image plane is performed. The extrema concerning 2D image coordinates (outer contours) are computed from the projected robot model and verified (matched) with the observation from the camera image. Corresponding regions are then considered to be a part of the robot (see Figure 3).

2 Attention Attractors

Before passing information to the *Analysis and Interpretation* stage described above, we apply a novel early processing mechanism, which is the main contribution of this paper. This section herein describes the effects directly triggered by the input data, whereas Section 3 explains the implemented cognitive feedback mechanism.

The goal of attention attraction described here is to generate a saliency map, basically a virtual image that indicates certain regions as being relevant for further analysis. As we want to extend the idea of the saliency map to a more general map of visual attention in the next section, we introduce the term *attention condensation layer*, also to emphasize our technical perspective.

As described in the introduction, we rely on two properties of our input data in order to extract and propagate relevant regions of attention. First, we try to find salient local features by analyzing intensity and hue, like e.g. [Hu *et al.*, 2008]. Second, we propose to analyze the dynamics we observe from a sequence of input images to extract further cues on regions that might be interesting.

Both algorithms described in this section directly process sensory input data, so we call the emerging effects *bottom-up* attention attractors. [Itti *et al.*, 1998] refer to these effects as "scene-dependent", on the contray to "task-dependent" ones that originate from higher cognitive processing (see Section 3).

2.1 Static Saliency

Our approach for detecting salient local features in a single, thus static, input image relies on a comparison of intensity and hue. A background model is used, which can be trained in advance. The model is built from a 2D normalized joint histogram [Pass and Zabih, 1999] representing the joint distribution of background pixel values with respect to their intensity and hue. Creating the 2D model is straight forward, as it is sufficient to only analyze one empty input frame, i.e. one that does not contain any objects, gestures or robot parts.

In the saliency detection step the model is compared to hue-intensity distribution of image patches in the input image. Here, computing e.g. the *Bhattacharyya* distance [Bhattacharyya, 1943] gives a measure of similarity. The lower the distance, the more similiar is a patch to the background and the less salient is the region. If the distance is greater than a certain threshold, the patch is considered to be worth analyzing it within the recognition stage (see Figure 5). As OpenCV already provides optimized algorithms and data structures for multi-dimensional histogram comparisons, this step can be integrated into the proposed vision system efficiently.

2.2 Dynamic Saliency

The extraction of saliency from dynamics we describe here, is an extension to the approach for detecting locally salient regions explained above. The basic idea is the evaluation of the object movements in a sequence of sensed input data. There are several algorithms for movement estimation (e.g. optical flow) in an image sequence considering different motion models (e.g. *Brownian motion*) and temporal levels of depth (history). The

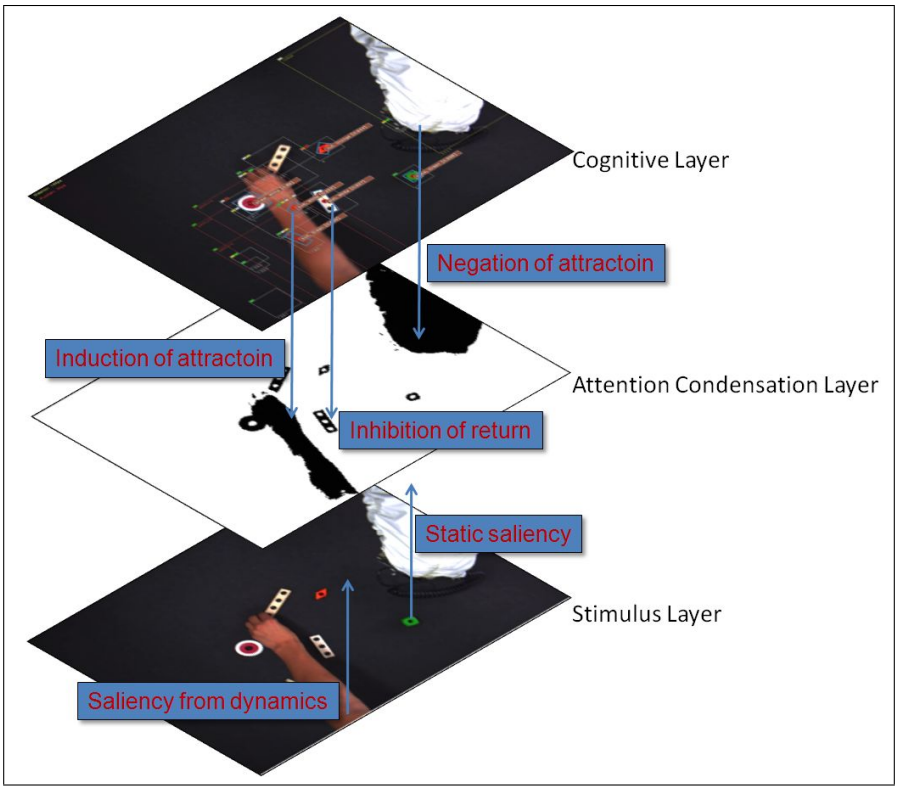

Figure 5: The visual layers for attention based information filter. The condensation mechanism evaluates influences from the cognitive layer and bottom-up attractors.

approach used in the proposed system is utilizing of the most common and straight forward: a *disparity map*.

In principle, we create and evaluate a disparity map, combine it with the saliency map and observe its behavior for a number of frames. A disparity map is a binary image where pixels are set to `true` if their value changes significantly within in the transition from $t^{n-\delta}$ to t^n, where δ is the number of timesteps back in history that influence the result, and `false` otherwise.

Regions of particular interest are blobs of pixels moving in an uniform manner for a number of frames δ. On the basis of this observation, the system is able to infer regions containing high dynamics, that are then considerd to be worth analyzing (see Figure 5). Anticipating the next section, this mechanism is the counterpart to the inhibition of return described there.

3 Cognitive Feedback

The second principle in our attention based robot vision system is the assumption, that the cognitive layer should be able to influence the effective level of attention payed to a bottom-up attracted region by giving some sort of feedback. Technically speaking, this means projecting knowledge about a scene or about constraints in the world into the attention condensation layer introduced in the last section. The active or sometimes even conscious projection of world knowledge can either cause inhibitory (Section 3.1) or creational / reinforcing (Section 3.2) effects.

Neuroscientists often call these effects on the primary visual cortex of humans *top-down* effects, as they have their origin on higher levels of cognition (e.g. [Li *et al.*, 2004]). Their experiments show, that the same bottom-up stimuli or as we called them before, attention attractors, have very different influence on the focus of attention and thus the activation of processing units under variations of the task to accomplish. [Itti *et al.*, 1998] call these effects "task-dependent" for the very same reason, i.e. as the high level task or plan influences lower level visual attention to specific regions.

The mechanisms and developed algorithms described below exploit the two different possibilities of influence, inhibition or reinforcement, on the attention condensation layer.

3.1 Inhibition of Return

Inhibition of return, a well known expression from psychology (see e.g. [Posner and Cohen, 1984; Posner *et al.*, 1985]), constitutes the theoretical foundation of one of the algorithms used to control the focus of attention in a top-down manner.

Concerning the JAST robot setup situations are to be considered, where the system's attention attractors were activated and regions for analysis from the bottom-up view were in turn identified. In this case the inhibition of return mechanism avoids reanalyzing regions that have been previously processed (Figure 5).

In order to achieve this, the system keeps track of any object, gesture or part of the robot visible in the scene. Many of these items are likely to appear at the same or very close position in consecutive frames. The level of attraction for a ROI $a(ROI, t)$ is in this case proportionally decreased with the number of sequential frames t it appears in. Here we propose two methods, either linear or non-linear degression.

$$a(ROI,t) = \begin{cases} \text{linear:} & \delta(ROI) - t \\ \text{non-linear:} & tanh(-\alpha t) + 1 \end{cases} \quad (2)$$

Within the linear degression, we have to specify a $\delta(ROI)$ which specifies the maximum number of consecutive attractions subject to the size of a region (larger regions need more time for analysis, therefore the inhibition of return affects big regions later), whereas within the non-linear case, we have to specify a factor α that determines the duration of the attraction and its strength subject to t. Usally $\alpha \simeq \delta(ROI)^{-1}$ is a good choice here.

Complete inhibition after a certain amount of time can be computed with respect to a signum thresholding function $T_\epsilon(ROI, t)$. In the linear case $\epsilon \geq 0$ is permitted, while $\epsilon > 0$ is a constraint in the non-linear case.

$$T_\epsilon(ROI,t) = \begin{cases} 1 & \text{if } a(ROI,t) \geq \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Considering computational effort, no more processing is performed and no more ressources are used for the analysis of a region if $T_\epsilon(ROI, t)$ evaluates to 0.

3.2 Volitional Attention Control

Thinking about human attention control again, we find it obvious and intuitivly clear that we are able to control our focus of attention and direct it to a certain region, or generally to a subset of the perceived input data. This was shown to be plausible in historical psychological experiments by [Stroop, 1935], and is still being researched on, e.g. [Cohen *et al.*, 2004].

There are basically two complementary options for this kind of attention control effects. First, if we suppose an object to be at a certain location, we are able to take a closer look, even when lacking the bottom-up attraction of this location. And second, even if there is bottom-up attraction, we can volitionally choose not to pay attention to this specific attractor and ignore the stimulus. The system proposed in this paper applies both control strategies, as it provides specific interfaces to the higher level cognition modules (indicated by top-down arrows in Figure 5).

Conscious Induction of Attraction on the one hand is an interface allowing for emphazising of specific or creation of new areas of interest in the attention condensation layer. Hence attention on existing regions is reinforced or regions of interest are generated artificially.

Negation of Attraction on the other hand, unlike the inhibition of return, is not automatically triggerd by the visual activation tracking mechanism, but rather originating from a higher level of cognition. Still, invoking the interface causes a similar effect: the computational effort put into the analysis of the designated region is repealed.

4 Priorization of Attention Mechanisms

There is an implicit order of the attention filters described in Section 2 and Section 3. From bottom to top the priorities are intuitively set as follows:

- Static saliency is the lowest priority attention attraction bottom-up mechanism. It is computed directly from a static input image.
- Dynamic saliency is the second lower priority bottom-up mechanism. As it already utilizes static saliency features for the evaluation, it has higher priority than static saliency.
- Inhibition of return is the lowest of the cognitive feedback mechanisms. Inhibition of return can inhibit attentional focus on static saliency features and slow dynamic features. As fast dynamics are allowed to overwrite inhibition of return, it can be considered to have the same priority than dynamic bottom-up features.
- Cognitive inhibition and cognitive reinforcement / creation of attionional focus to a region is the highest priority mechanism in the current version of the system. The mechanism overwrites all attention selections computed in the lower priority modules.

The system proposed in this paper applies this priority ranking in each cycle of the preprocessing stage.

5 Experimental Results and Conclusion

The vision system described in this paper has been evaluated with respect to the benefit of the improvements with the proposed attentional mechanisms. Hence we created a test-bed where it is possible to switch the described algorithms of Section 2 and Section 3 on or off.

In order to obtain meaningful results, we capture different input videos of typical interaction scenarios in advance and use these video streams to feed our vision system with the algorithms enabled or disabled. The videos are recorded at a sampling rate of 7 fps with a

resolution of 1024×768 pixels. Our test system is standard PC hardware, equipped with an AMD Athlon™, 64 X2 Dual Core Processor 3800+ and 2 GB RAM.

Figure 6 exemplarily depicts an image of a dynamic input sequence[2] and the analysis performed. In the figure some relevant information for high level cognitive / reasoning modules is annotated: a gesture is recognized and several objects are detected and their location, orientation and color is identified. The boxes around the identified items are not published, but indicate the corresponding ROIs extracted with the attention attraction algorithms.

In Figure 6 one can see all of the effects described in the above sections. The secene is taken from a typical input video sequence, where a subject has just placed a slat on the table, while the robot was standing still.

Static saliency Region 1421, the slat in the subject's hand, is computed with this approach. Its intensity differs a lot from the background and so the previously unseen region is considered to be worth paying attention.

Dynamic saliency The red regions, likely to be gesture regions, are extracted using the moving blob approach. In the image, the last few blobs are depicted. The regions are likely to be gesture regions because the blob was moving into the scene from the bottom, which is a presumable position for a subject in the JAST setup. Also, regions 1391–1393 in the image originate from scene dynamics, but unfortunately they are false positives.

Inhibition of return Regions 4 and 5 containing an orange nut and a green cube are not re-analyzed, although there is a static stimulus and attention attractors were activated by the bottom-up mechanism. In this case the inhibition of return mechanism avoids the waste of resources on these specific regions. They have very low ids, which indicates that they have already been tracked for many frames.

Conscious induction of attraction Regions 1185 and 1178, the virtual objects, are projected by the cognitive layer. From the whole video sequence one can see, that these objects are actually lying on the table, but cannot be seen by the system in the snapshot-image. The cognitive layer therefore infers, that objects apparently do not disappear from a scene so suddenly. Thus a virtual region of interest is generated for these formerly visible objects and the regions are reanalyzed.

Negation of attraction For region 1357, the small one next to the robot, the cognitive layer prohibits further analysis, as it is part of the inferred position of the robot and thus does not contain relevant information. Although, by means of the region size it could possibly contain an object, the poperty "no object" is assigned.

[2]http://www.youtube.com/watch?v=JupXjgdYzY4

One has to consider, that a performance analysis cannot be performed straight forward, because, as shown in [Müller *et al.*, 2008], the system operates massively parallel in the function **and** the data domain. This means, lack of computational ressources are compensated for with frame-drops. But still, we are able to measure according to a very basic metric: the time it takes, until each object in a scene is detected and analyzed.

In order to have reliable ground truth data, we first analyze the system's performance without any attention driven improvements. Quickly we find, that evaluation even for a system only using object recognition (no gesture and no robot recognition) with a template matching approach of one single image with respect to 16 possible template objects (a typical number for the JAST setup) and 20 rotations takes more time than disposeable in the application scenario. The naive implementation takes around 120 seconds to analyze a single frame containing 21 objects! But, when applying the bottom-up attention attraction mechanism of static saliency only, the amount of time needed for processing the whole scene already decreases to 6.39 seconds. This denotes an improvement with of a factor 18.78.

Taking a video input sequence to evaluate the improvements based on image dynamics, we see, that due to inhibition of return, the capability of **tracking** regions, the computation times again decrease. As most of the objects on the table in the JAST setup are static as long as neither the robot nor the human moves, this mecahnism is an efficient way to lower the complexity. Consider the frame from above, the analysis took 6.39 seconds, but enabling the inhibition mechanism improves the performance to more than real-time, once all regions were analyzed.

In order to show the value of attraction on dynamics and conscious attention focussing, we consider the example of moving a hand or robot arm in the scene. First, dynamic saliency compensates for inhibition of return, so moving objects are reanalyzed although the stimulus itself might remain almost static, but spatial changes or distortions trigger the analysis. Second, conscious mechanisms allow to compensate for illogic attention attraction, such as unnatural region behaviour (sudden disappearances or appearances) or false positives due to errorneous saliency.

Summarizing our work, we propose a biologically inspired robot vision system for a human-robot interaction scenario. The target of the project is to build a system capable of natural, and thus quick, actions and reactions. Hence, we apply theoretical findings on human visual apparatus to a technical system in a slightly simplified way. Therewith we show, that the implementation not only improves the performance of the overall

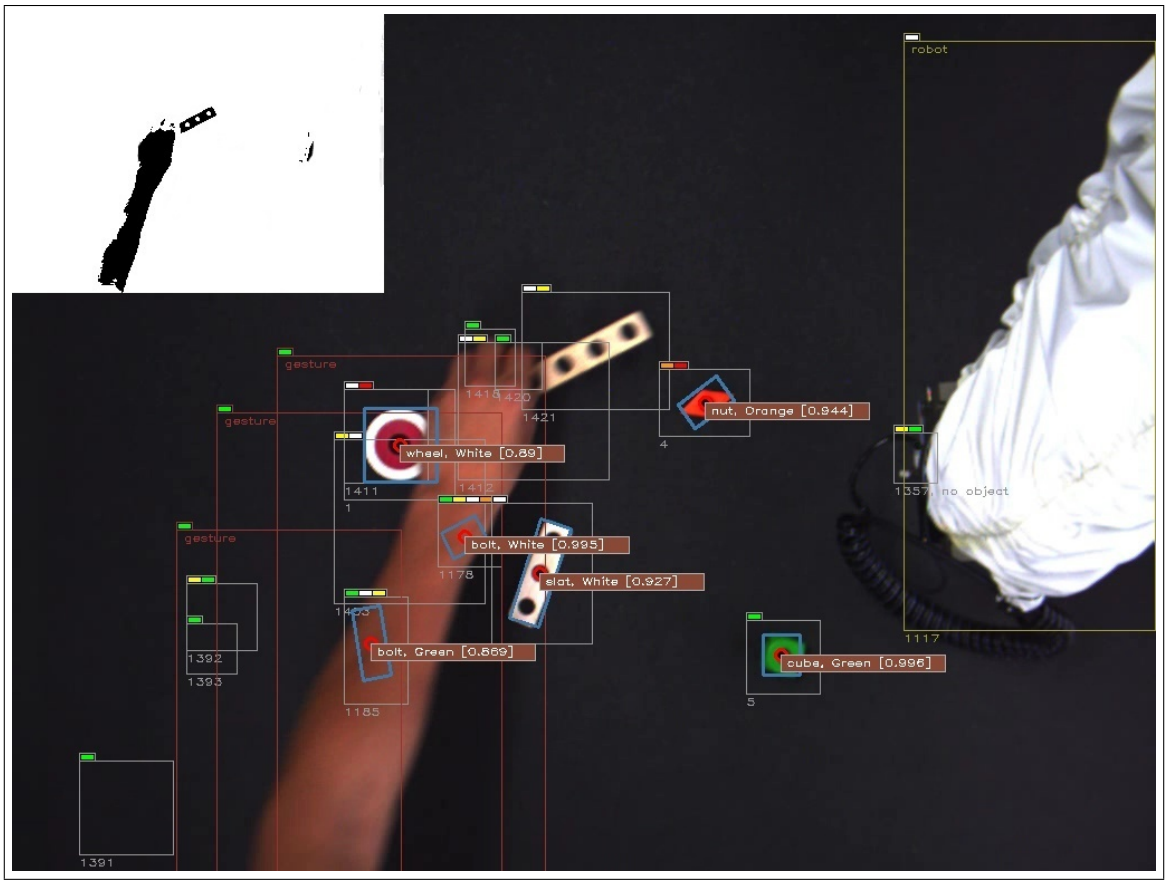

Figure 6: On analysis, the system is supposed to publish some of the information annotated in the figure. High-level information are position, orientation, color, classification and certainty for objects and the robot's and gestures' most likely locations. Regions of interest, projected world knowledge (induced attraction, negation of attraction), dynamics and inhibition of return are meta information. The figure also shows a snapshot of the binarized attention condensation layer in the upper-left corner.

system, but also mimics biological systems and indicates the plausiblity of previous theoretic work.

6 Acknowledgements

This research was supported by the EU integrating project JAST "Joint-Action Science and Technology" (FP6-003747-IP). Please visit http://www.euprojects-jast.net for further information.

References

[Bhattacharyya, 1943] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[Cohen *et al.*, 2004] Jonathan D. Cohen, Gray Aston-Jones, and Mark S. Gilzenrat. A systems-level perspective on attention and cognitive control: Guided activation, adaptive gating, conflict monitoring, and exploitation versus exploration. In Michael I. Posner, editor, *Cognitive Neuroscience of Attention*, chapter Cognitive Models of Attention, pages 71–90. Guilford Publications, 2004.

[Culler *et al.*, 1999] David E. Culler, Jaswinder P. Singh, and Anoop Gupta. *Parallel Computer Architecture: A Hardware/Software Approach*. Morgan Kaufmann Publishers, 1999.

[Foster *et al.*, 2007] Mary Ellen Foster, Thomas By, Markus Rickert, and Alois Knoll. Humanrobot dialogue for joint construction tasks. In *Proc. ICMI*, 2007.

[Gilles, 1998] S. Gilles. *Robust Description and Matching of Images*. PhD thesis, University of Oxford, 1998.

[Giuliani and Knoll, 2007] Manuel Giuliani and Alois Knoll. Integrating multimodal cues using grammar based models. In *Proc. ICMI*, 2007.

[Herlihy, 1991] Maurice Herlihy. Wait-free synchronization. *ACM Transactions on Programming Languages*

and Systems (TOPLAS), 13(1):124–149, 1991.

[Hu *et al.*, 2008] Yiqun Hu, Deepu Rajan, and Liang-Tien Chia. Detection of visual attention regions in images using robust subspace analysis. *Journal of Visual Communication and Image Representation*, 10(3):199–216, April 2008.

[Itti and Koch, 2001] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, March 2001.

[Itti *et al.*, 1998] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

[Kadir and Brady, 2000] Timor Kadir and Michael Brady. Saliency, scale and image description. *IJCV*, 2000.

[Kienzle *et al.*, 2006] Wolf Kienzle, Felix A. Wichmann, Bernhard Schölkopf, and Matthias O. Franz. A non-parametric approach to bottom-up visual saliency. In *Proc. NIPS '06*, 2006.

[Li *et al.*, 2004] Wu Li, Valentin Piëch, and Charles D. Gilbert. Perceptual learning and top-down influences in primary visual cortex. *Nature Neuroscience*, 7:651–657, May 2004.

[Message Passing Interface Forum, 1995] Message Passing Interface Forum. MPI, A Message-Passing Interface Standard. Technical report, University of Tennessee, Knoxville, Tennessee, June 1995.

[Müller *et al.*, 2008] Thomas Müller, Pujan Ziaie, and Alois Knoll. A wait-free realtime system for optimal distribution of vision tasks on multicore architectures. In *Proc. ICINCO '08*, 2008.

[ope,] Open Source Computer Vision Library (OpenCV). Technical report, Intel Corporation.

[O'Regan and Noë, 2001] J. Kevin O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioural and Brain Sciences*, 24:939–1031, 2001.

[Pass and Zabih, 1999] Greg Pass and Ramin Zabih. Comparing images using joint histograms. *Multimedia Systems*, 7(3), 1999.

[Posner and Cohen, 1984] M.I. Posner and Y. Cohen. Components of visual orienting. *Attention and Performance*, 10:531–556, 1984.

[Posner *et al.*, 1985] Michael I. Posner, Robert D. Rafal, Lisa S. Choate, and Jonathan Vaughan. Inhibition of return: Neural basis function. *Cognitive Neuropsychology*, 2(3):211–228, August 1985.

[Reinagel and Zador, 1999] P. Reinagel and A. M. Zador. Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems*, 10(4):341–350, 1999.

[Rickert *et al.*, 2007] Markus Rickert, Mary Ellen Foster, Manuel Giuliani, Thomas By, Giorgio Panin, and Alois Knoll. Integrating language, vision, and action for human robot dialog systems. In *Proc. ICMI*, 2007.

[Stroop, 1935] John Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Pyschology*, 18:643–662, 1935.

[Sundell and Tsigas, 2003] Hakan Sundell and Philippas Tsigas. Fast and lock-free concurrent priority queues for multi-thread systems. In *Int. Parallel and Distributed Proc. Symp.*, 2003.

[Walther *et al.*, 2002] Dirk Walther, Laurent Itti, Maximilian Riesenhuber, Tomaso Poggio, and Christof Koch. Attentional selection for object recognition – a gentle way. In *Biologically Motivated Computer Vision*, pages 251–267, 2002.

[Ziaie *et al.*, 2008a] Pujan Ziaie, Thomas Müller, Mary Ellen Foster, and Alois Knoll. Using a naïve bayes classifier based on k-nearest neighbors with distance weighting for static hand-gesture recognition in a human-robot dialog system. In *Proc. Intl. CSI Computer Conference '08*, 2008.

[Ziaie *et al.*, 2008b] Pujan Ziaie, Thomas Müller, and Alois Knoll. A novel approach to hand-gesture recognition in a human-robot dialog system. In *Proc. of the First International Workshop on Image Processing Theory, Tools and Applications*, 2008.