Scenes and Tracking with Dynamic Neural Fields: How to Update a Robotic Scene Representation

Stephan K. U. Zibner∗, Christian Faubel∗, Ioannis Iossifidis∗, Gregor Schöner∗, and John P. Spencer†

∗ Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
Email: (zibner|faubel|iossifidis|schöner)@neuroinformatik.rub.de

† Department of Psychology, University of Iowa, USA
Email: john-spencer@uiowa.edu

Abstract—We present an architecture based on the Dynamic Field Theory for the problem of scene representation. At the core of this architecture are three-dimensional neural fields linking feature to spatial information. These three-dimensional fields are coupled to lower-dimensional fields that provide both a close link to the sensory surface and a close link to motor behavior. We highlight the updating mechanism of this architecture, both when a single object is selected and followed by the robot’s head in smooth pursuit and in multi-item tracking when several items move simultaneously.

I. INTRODUCTION

Autonomous robots designed to interact with human users need to build up and maintain a scene representation based on their own sensory information. Such a scene representation enables a robot to respond efficiently to user commands that refer to spatial locations, to object features, or to object labels, without having to perform a visual search each time a command is interpreted.

The extraction of meaningful information about the robot’s environment through perceptual systems is currently one of the major bottlenecks that holds back the development of autonomous robots. For mobile robots, self-localization and mapping (SLAM), is a more elementary problem, towards which much progress has been made over the last decades [1]. To generate goal-directed action that goes beyond moving to a particular location, robots need to have extended maps, in which objects are segmented [2], [3], and identified [4], [5]. To enable the reaching and grasping of objects, such a representation needs to include pose information about objects [6]. All three aspects of segmentation, identification, and pose estimation are currently underdeveloped. Another aspect of scene representation for robots is that objects [7] or object categories [8] must be learned on the fly from a small number of exposures.

Here we propose a biologically inspired architecture for scene representation in a concrete scenario in which workspace (a table) is shared between our robotic assistant CoRA [9] and a human user. The architecture sequentially builds up a neuronally based representation of objects contained in the shared workspace, the scene. Access to this representation through cued recall task [2] is also realized. A scene representation must provide the robot with long-term memory for object labels which it must be able to link to object features and to the spatial configuration. Because the interaction with human users takes place under dynamic conditions in which objects may be occluded or leave the viewing range, a scene representation must be endowed with working memory. To maintain a match with the scene, a scene representation must be continuously updatable. When, for instance, a user moves objects within the workspace, the scene representation must autonomously update its spatial representation to keep the links between spatial and feature information consistent.

Here, we focus on different forms of object tracking that are one form of how scene representations may adapt in changing environments. Tracking is a problem well-known in computer vision (see [10] for a review of visual tracking algorithms), and supports the detection and classification of specific categories of objects such as in pedestrian or car tracking in visual traffic scenes [11]. Visual tracking is not considered as often in the context of scene representation.

Two different forms of tracking relevant to scene representation may be distinguished. One is the tracking of a single object that is in the focus of attention. Active vision, in which a moving camera system may keep the moving object within the visual array may be part of such tracking, which makes it possible to track over a larger portion of the visual array.

PREPRINT
than the field of view of a camera would allow. An example of such an active vision systems is [12]. The second type of tracking keeps the representations of multiple objects linked to their changing positions in parallel. Such multi-item tracking has been addressed in the framework of SLAM to keep the representation consistent [13], [14].

Humans perform both types of tracking. When a single object is tracked, attention is typically directed to the object. The eyes may follow a selected object in a smooth pursuit movement that is stable in the presence of competing distractors [15]. When an object moves too fast, the same visual error signal suberves catch-up saccades that reinstate the tracking state [16]. Humans are also capable of tracking multiple objects, typically up to four at a time, even in the presence of distractors or when targets are temporarily occluded [17], [18]. The task increases in difficulty, if feature information like object color must be tracked as well, which requires that the object features must remain bound to the spatial location [19].

Here we present a Dynamic Neural Field Architecture that implements both types of tracking. At the core of the architecture is a three-dimensional field that represents the two-dimensional spatial location of objects on a surface together with their color. This representation is built up sequentially, one object being brought into foreground after another. When an object is selected, its feature value — color — is associated with its spatial position and that association is represented in the three-dimensional space-color field. Color as a simple feature dimension stands for potentially more complex feature vectors. These may become linked to object labels in an object recognition system, a process that typically requires bringing a single object into the foreground (for a system like this that uses the same theoretical language see [20]). If the object moves while it is in the perceptual foreground, then the first form of tracking by active camera movement (smooth pursuit) is induced. If one or multiple objects move after a space-color association has been built across a set of objects, then multi-object tracking takes place and keeps track of updated items. While attention is not redirected to an individual object, only its spatial location, not its feature values, may be updated. During this form of multi-item spatial tracking, the associated color values need to remain linked to each tracked object.

II. SCENE REPRESENTATION ARCHITECTURE

The architecture for scene representation consists of ten Dynamic Neural Fields that are coupled in complex ways. Here we only discuss those layers that contribute to the two forms of tracking. The full architecture and supported functions other than tracking are described in [?].

A. Dynamic Neural Fields

Conceptually, Dynamic Neural Fields are dynamical neuronal networks, in which the discrete sampling of relevant perceptual or motor dimensions by individual neurons is replaced by continuous distributions of neuronal activation (for a conceptual introduction, see [21]). The evolution of the field's activation variable $u(x, t)$ is defined over a continuous metric parameter $x$ such as the spatial position of an object is captured by the following dynamical equation [22]:

$$
\tau_1 \dot{u}(x, t) = -u(x, t) + h + s(x, y, t) + \int w(x - x') \sigma(u(x', t)) dx' \tag{1}
$$

Without input the field relaxes to its resting level defined by $h$, with input the field may locally pass threshold and build a peak defined through the field interaction, which is expressed by the interaction kernel $w(x - x')$ and the threshold function $\sigma(u(x', t))$. Localized peaks of activation are units of representation. When the activation level in the peaks exceeds a threshold (conventionally chosen to be zero), such peaks represent perceptual or motor decisions, both in the sense of detection and in the sense of selection among competing inputs. The location of such peaks along the feature or motor dimension represents a metric estimate of the perceptual or motor state. Such peaks are attractor states of the neuronal dynamics, that may coexist bistably with sub-threshold distributions of activation and may go through instabilities. Depending on the parameter values, Dynamic Neural Fields may be set up for doing selection decisions when presented with multiple localized inputs or, alternatively, may support multi-peak attractor solutions [23]. Again depending on parameter values, peaks may either be self-stabilized by neural interaction but still dependent on input, or self-sustained by neuronal interaction alone [21]. The latter class of attractor solutions represents a form of working memory. Long-term memory is implemented by slowly evolving memory traces that are laid down in a field wherever activation is above threshold. Memory traces couple back into a field as excitatory input and thus preshape the field.

Outside instabilities, single- or multi-peak solutions are stable states. Therefore, the perceptual or motor decisions that such peak solutions represent persist when input fluctuates or is briefly removed. Nevertheless, peak solutions continue to be sensitive to changes in the environment. As a distribution of localized inputs moves around, its maxima may be tracked by the peaks of a Dynamic Neural Field. Such multi-item tracking has first been demonstrated by [24].

B. The Different Levels

We structure the system into four different levels that have different degrees of invariance and may be associated with different areas of the human cortex. The retinal level is closest to the sensory surface with the lowest degree of invariance and could be viewed as a functional description of visual cortex. At the scene level, spatial information is represented together with object feature information in an allocentric reference frame that is attached to the shared workspace (the table). This level may be associated with the lateral intraparietal area (LIP) [25]. The infero-temporal cortex is associated with visual object representation [26], which in our architecture happens at the object level. Finally, at the motor level head motion is represented, thus closing the action-perception loop. This level
may be neuronally associated with the frontal eye fields [27], the superior colliculus [28], and the brain stem [29].

C. Retinal Level

Segmentation, selective attention, and low level feature extraction take place at the retinal level. At this level, sensory input is highly variant, the sensory stream changes when objects move in the environment, but also with every head movement. Three Dynamic Neural Fields are at work at the retinal level.

1) Visual preprocessing: The first stage of visual input computes a simple saliency map by calculating on- and off-center responses on image intensity and on two opponent color channels. At each retinal location, the three responses are summed with equal weights into a single saliency value. This simplified version of the salience computation of [30] is sufficient in our scenario. In addition to this saliency map we compute a hue color map that serves as input to the retinal color field.

2) Retinal Space Field: The result of the saliency computation is directly mapped onto a two-dimensional neural field that supports multiple peaks and represents object locations. The output of this field is a normalized and stabilized version of the saliency map that is fed into the retinal selection field. A second projection goes to the scene space field on the scene level after a spatial transformation into the allocentric table frame that uses the known head orientation. Here this mapping is computed algorithmically but such mappings may well be learned by neural systems [31].

3) Retinal Selection Field: A single spatial location is brought into foreground to extract feature information. This happens in the retinal selection field, which receives input from the retinal space field and operates in single peak mode. The selected peak is used to compute a feature histogram local to the peak location. The output of the retinal selection field is projected into the scene space-color field at the scene level. This projection provides tube input, which is localized in two-dimensional space but constant along the feature dimension. The retinal space selection field also projects to the motor selection field so that the selected item is centered on the camera plane.

4) Retinal Color Field: Extracting feature values that characterize a possible object at a particular retinal position is achieved with the aid of the retinal space selection field. The peak in this field is used to mask all retinal regions except within the peak. All visual locations with supra-threshold activation in the selection field pass their hue values into a color hue histogram. This histogram is used as an input distribution in a feature field defined over hue. A detection instability induces one or multiple localized peaks that represent the dominant object colors within the selected spatial region. The field output is fed as a ridge input into the label-color field (constant along the label, localized along the feature dimension) at the object level and as slice input into the scene space-color field (localized along the feature dimension, constant along the two spatial dimensions) at the scene level.

D. Scene Level

At the scene level, spatial position is represented in an allocentric reference frame attached to the table. This representation is thus invariant to head movements. Because it receives input from the retinal level, it is still able to track moving objects, however. Two different fields are at work at this level, the scene space field that represents only the spatial configuration of the scene, and the scene space-color field that represents the object feature color over two-dimensional space.

1) Scene Space Field: The scene space field supports multiple attractors and provides the system with spatial working memory through self-sustained peaks for object locations that are out of sight. It also supports self-stabilized peaks within the current field of view that keep track of visible objects. The field receives input from the visual pre-processing and the retinal space field that are transformed to the allocentric table frame.

2) Scene Space-Color Field: The key component of the scene representation is this three-dimensional neural field. This field provides working memory for associations between one-dimensional color information and two-dimensional spatial information. These associations are memorized sequentially: Every time an object is brought into foreground, a new working memory peak is created. Despite this sequential model of creating memory items, the field continuously tracks spatial changes in the scene. When an object is removed from the scene, the associated working memory vanishes.

This functionality is implemented through continuous coupling to the scene space field, which provides tube input specifying the spatial locations of objects. Note that objects outside of the viewing angle cannot be updated. Thus, if an object is moved while it is outside the view, it appears as a new object when it returns into view. The old association is
Fig. 3. This figure shows an overview of the architecture for robotic scene representation. For simplicity, space is reduced to one dimension. Transfer of activation between fields is illustrated by either solid or dotted arrows, where the latter ones depict an included reference frame transformation. Field activity is represented by solid red lines. Blue dotted lines are excitatory input and green dashed and dotted lines are inhibitory input. One of four possible objects in this sketch is already scanned and stored in the associative scene space-color field E. The three leftmost objects are contained in the current retinal space field A as well as in the scene space field D. Both are mutually coupled and receive additional input from visual preprocessing. The scanned object is inhibited in the retinal space selection field B, which receives input from A. The still-active selection triggers a localized color extraction, which is represented in the retinal color field C. B is also coupled to a motor selection field G, which in turn is linked to the motor system. B, C, and D are also coupled to E. The recall of stored object locations and features is done in the scene space-color selection field F. The label-color field H contains long-term memory associations between object labels and color hue. Features are provided by the retinal color field C, whereas labels are given by the user. 

removed because the tube input that sustains this association vanished.

When a working memory peak for an association is first created, three fields contribute their outputs. In addition to the spatial tube input from the scene space field, the retinal selection field provides a single item tube input that boosts a single spatial location within the scene space-color field boost. Only at that location is the slice input along the color dimension from retinal color field capable of building a new working memory peak.

E. Object Level

At the object level, the feature representation depends neither on spatial changes within the scene nor on head movement. The representation at this level has the highest degree of invariance. Feature input from the scene is only provided when an object has been selected in the retinal selection field.

1) Object Label-Color Field: The object label-color field is a two dimensional association field representing the hue color along one dimension and discrete labels along the other dimension. The feature input comes from the retinal color field. When users provide label information, the label-color field receives ridge input along the feature dimension at the specified label. Where this input intersects with the ridge input along the label dimension from the retinal color field, a peak builds. That peak drives to the build-up of a long-term memory trace that represents the label-feature association. Once this memory trace has been created, it preshapes the association field. When feature input is now received, that matches the learned feature-label association, a peak is built without the need for a user to specify the label.

F. Motor Level

At the motor level head movements are planned in angular coordinates and the associated motor signals are generated. The motor selection field receives input from the retinal selection field and projects onto two separate one-dimensional fields at higher angular resolution, the motor pan field and the motor tilt field, that represent the pan and the tilt angles of the camera.

G. Sequence Generation

The sequential organization of activation states that leads to the various operations of memory building and updating
emerges from the dynamics of neural interaction. A peak in the label-color field represents that an association has been learned. That is the signal for bringing a new object into the foreground. Therefore, whenever a peak is formed in the label-color field, a peak detector function sends a negative boost to the selection field which extinguishes any peak there. The selection field also receives inhibitory localized input from the space-color representation at the scene level, which effectively reduces the propensity for a peak to build at a spatial location that has already been in the foreground earlier and registered an item in working memory at the scene level.

III. RESULTS

The implementation of the architecture for robotic scene representation was deployed on the robotic platform CoRA. We designed four different experiments in order to evaluate the system’s ability to track and to update and to evaluate the system’s limits of memory capacity. To test those in a systematic and reproducible way, we use small mobile robots (E-Puck\(^1\)) with color markers, which are placed on the workspace of our robotic assistant CoRA. They are set to run in Braitenberg [32] obstacle avoidance mode, which is already built-in as a default behavior. In the two experiments III-B and III-C that evaluate the multi-item tracking capability and the capacity limits, CoRA's head is fixed and the area in which the robots may move is restricted to CoRA's field of view. Without this limitation it is not possible to assure for a constant number of robots within the field of view. How the system can cope with robots leaving and entering the field of view is demonstrated in experiment III-D.

A. Experiment 1 - Active Selection and Smooth Pursuit Tracking

The first experiment is a test of single-item smooth pursuit tracking. CoRA is confronted with a scene consisting of a single colored mobile robot and two distractors. When it selects the moving robot it is kept in the state of keeping this robot in the foreground.

Results: Figure 4 illustrates the stability of such a once-made selection decision when a visual distractor comes close to the selected object. At the same time, this run exhibits tracking of the selected object through a pursuit movement of the camera head. Selection is stable even for moving objects. The distracting obstacles did not influence the color extraction for the selected object. Scene information was accurately updated at all times through the tracking peak in the three-dimensional scene space-color field.

B. Experiment 2 - Multi-Item Tracking

This experiment demonstrates the second form of tracking and updating a scene. For this experiment, CoRA's head was fixed and the visible area of the table was surrounded by walls to keep the moving objects (colored robot vehicles) within viewing range. The moving robots are registered in the scene space-color field on the fly while moving. The task is to maintain the space-color associations that identify the robots while updating location information simultaneously for multiple robots.

Results: Multi-Item Tracking was tested with two and three mobile robots. A timeline of a multi-item tracking course can be seen in Figure 5. While two robots could be tracked with ease, adding a third robot drove the architecture to the edge of stability so that slight perturbations may lead to the loss of a working memory peak in the scene space-color field and therefore the loss of the binding of that object’s position and its feature information. The architecture is capable of multi-item tracking, but has capacity limits much like humans do.

C. Experiment 3 - Capacity Restrictions And Tracking

As a follow up we examine more precisely the capacity limits of the system. A number of objects, including one mobile robot, are placed on the table. The architecture then sequentially registers the objects through working memory peaks in the scene space-color field. The number of already registered objects determines, if another object can be included in the scene representation [23]. After reaching the capacity of the working memory, the mobile robot is activated and starts moving in the scene. That change in task may alter the effective capacity of the system.

\(^1\)http://www.e-puck.org/
Results: The robot was confronted with a scene consisting of four objects (see Figure 6), which the robot was capable of registering in the scene space-color field. A fifth object was temporarily added to the scene, but could not be included in the scene representation, once all four objects were registered, establishing static working memory capacity at four. After powering up the mobile robot, the movement in the scene led to the loss of one space-color association. That association was the one linking the robot’s position to the robot’s color. The change in capacity illustrates how tracking increasing the demands of the neural scene representation.

Fig. 6. Working Memory Limits. The displayed scene contains four objects of which only a single one can move around. After scanning the scene and registering the objects with space-color associations, the robot starts to move around. Since the tracking of a peak slightly decreases its input strength and the scene space-color field is already at its capacity limit of multiple co-existing peaks, the space-color association dissolves.

D. Experiment 4 - Scene Updates and Long-Term Memory

The last experiment demonstrates how to cope with the limited capacity of working memory and how to keep scenes up to date even when known objects leave and re-enter the visual area. For this experiment, two colored mobile robots are used. Both are known to the system in the sense that a long-term memory trace in the object label-color field exists for both objects. In this setup, the selection and the subsequent color extraction of an object leads to a peak in the object label-color field, because the color information matches the deposited preshape. Since a peak in the object label-color field automatically cancels out the selection, the examined object is immediately released from the system’s explicit attention. The course of the experiment lets the red robot circle in the visual area of CoRA’s camera, while the blue robot starts somewhere outside the visual area on the table, but enters the visual range after a short period of time. The described mechanism of selection and recognition through feature matching in the object label-color field is a long-term memory driven mechanism to register the newly entering object in the scene representation.

Results: The course of this experiment can be reconstructed in Figure 7. The red robot was tracked by the scene color-space field at all times, providing a valid space-color association for this object. A short time after the blue robot entered the visual range of CoRA, a peak arose in the retinal space selection field, which lead to a feature extraction and a peak in the object label-color field marking a successful recognition of the robot. The selection was subsequently cancelled out, leaving the system in a state of multi-item tracking.

IV. DISCUSSION

A. Summary

We have presented a neurally inspired approach to robotic scene representation based on Dynamic Field Theory. The system’s ability of tracking and continuously updating the scene representation are emergent properties of the neuro-dynamic architecture. We demonstrated theses capabilities by confronting the system with typical changes in a scene, which required proper updates of the internal representation of space-color associations for all registered objects. Capacity limits, which also exist in the human counterpart of scene representation, were found for both static scenes and scenes in which objects were moving. With the existing implementation and computational restrictions, the capacity limit was four for static scenes and two to three for multi-item tracking. We demonstrated how long-term memory traces of scene objects help to register those objects on the fly, which reappear in a scene. This mechanism was an update of scene working memory for objects that did not have an active space-color binding in the scene representation.

B. Relation to Models of Smooth Pursuit and Saccadic Eye Movement

Other approaches to active perception of scenes [12] have focussed on controlling the switch between saccadic eye movement and smooth pursuit in a robotic head, inspired by an analogy with human eye movement. Our focus has been on the representation that emerge from visual control. In particular, we asked how to keep an internal representation up to date with the changing dynamic world. In addition to an attentional mechanism that is similar to smooth pursuit eye movement this also requires multi-item tracking capabilities.

C. Outlook

The color feature is essentially a placeholder to be replaced in future implementations by more complex object recognition systems. In addition to the ability to continuously track, a scene representation system needs a mechanism for change detection that will signal changes to objects that were out of sight. Linking to a DFT model of change detection in visual working memory [33] may enable the system to autonomously decide when to initiate an update of elements of its representation.

ACKNOWLEDGMENT

The authors acknowledge support from the German Federal Ministry of Education and Research within the National Network Computational Neuroscience - Bernstein Fokus: "Learning behavioral models: From human experiment to technical assistance", grant FKZ 01GQ0951.

PREPRINT
Fig. 7. Multi-Item Tracking and Long-Term Memory. This sequence of seven sequential steps shows the capability of the system to autonomously update the scene representation with the aid of stored long-term memory of label-color associations. The red robot is throughout present in the scene representation as shown in the upper field plots taken from a slice of the three-dimensional scene space-color field. The blue robot enters the scene in the second step and is selected. The stored long-term memory in the object label-color field is used to recognize the robot and to cancel out the active selection. After the recognition and registration, both robots are represented from the third step on and are tracked as shown in Section III-B.

REFERENCES


