Interactive Online Multimodal Association for Internal Concept Building in Humanoids

Christian Goerick, Jens Schmüdderich, Bram Bolder, Herbert Janßen, Michael Gienger, Achim Bendig,
Martin Heckmann, Tobias Rodemann, Holger Brandl, Xavier Domont, Inna Mikhailova
Honda Research Institute Europe GmbH
Carl-Legien-Strasse 30, 63073 Offenbach / Main, Germany
Email: christian.goerick@honda-ri.de

***Abstract*— In this paper we report the results of our research on learning and developing cognitive systems. The results are integrated into ALIS 3, our Autonomous Learning and Interacting System version 3 realized the humanoid robot ASIMO. The results presented address crucial issues in autonomously acquiring mental concepts in artifacts. The major contributions are the following: We researched distributed learning in various modalities in which the local learning decisions mutually support each other. Associations between the different modalities (speech, vision, behavior) are learnt online, thus addressing the issue of grounding semantics. The data from the different modalities is uniformly represented in a hybrid data representation for global decisions and local novelty detection. On the behavior generation side proximity sensor driven reflexive grasping and releasing have been integrated with a planning approach based on whole body motion control. The feasibility of the chosen approach is demonstrated in interactive experiments with the integrated system. The system interactively learns visually defined classes like "left", "right", "up", "down", "large", "small", learns corresponding auditory labels and creates associations linking the auditory labels to the visually defined classes or basic behaviors for building internal concepts.**

I. INTRODUCTION

The long term target of this research is a learning and developing cognitive system. Since this is a demanding target it can not be reached in one step but has to be approached as a long term research endeavor. This target is shared with several other researchers in the community of autonomous mental development as well as humanoid robotics, that has already lead to some interesting research activities and results.

Among others we would like to mention the research of the iCub project exemplarily presented in [1]. It addresses a complete humanoid infant including learning and mental development mechanisms. After setting up the robotics platform first experiments are being carried out. Some recent conceptually related work is presented in [2], where relations between affordances given as tuples of actions, object properties, effects and given words are learned offline. The approach presented in [3] also aims at a comprehensive system dealing with bindings between different modalities based on a working memory concept.

Roy [4] introduces the concept of schemata as behavior oriented scene description, bridging the gap between vision, speech, and behavior representations. Their system is able to discriminate between descriptive- and directive speech. Descriptive speech is understood and translated to memory updates as induced by perception. The interpretation of directive speech utterances leads to changes in the robots goal state, thus triggering the planning abilities to achieve the tutors wishing. The focus of this work is clearly on language understanding and not on learning. The lexicon and grammar of the speech recognition system are pre-coded, as well as the effects of e. g. the word "left" on the scene memory.

Iwahashi et al. [5] present a system which does not only learn the speech labels, but also the grammar. In a carefully controlled scenario, their system is able to learn words for objects, movements and concepts (like toys or tools). Due to the use of common belief propagation, the system requires a closed world assumption in which the a priori probabilities of words and objects are known. Furthermore the system can only learn from synchronously perceived visual and auditory stimuli.

We have chosen an iterative approach on the systems level in order to advance towards our target. This has already lead to a series of systems exemplifying the state of our research wrt. the current elements and the architecture hypothesis of cognitive systems [6], [7]. Our ongoing work is driven by a long term strategy and characterized by increasing functional performance under less constraints for the interaction and the internal learning processes. The systems are integrated on the humanoid robot ASIMO. For each research and integrated system state we aim at a complete system instance with sensory perceptions, an action repertoire, a global behavior organizing architecture and internal bootstrapping and learning mechanisms.

The system we present and evaluate within this paper is called ALIS 3. ALIS 3 is a rather comprehensive system, but here we will focus on the following scientific question: How to learn new perceptual classes like object properties and utterances and associate them as internal concepts during interactions with a robot? For answering this question we had to realize several crucial elements: The first is the local learning of sensory representations with possible top-down bias. As a result the system can learn new classes independently from each other without the strict need for co-occurrence. This does not imply that learning has to proceed sequentially: New classes as well as their association can be learned in parallel within the same interaction. That means the system can learn

a new object property like large object size, a corresponding utterance in any language and bind both together to the concept representing ”large”. A second major contribution is a hybrid data representation for supporting both local novelty detection and global classification and generalization. These representations facilitate the abstraction from locally defined examples into classes that can be globally evaluated without loosing the knowledge about which kind of data have already been experienced. This allows for bootstrapping concepts like ”left” and ”right” from a few examples in front of the robot but e.g. classifying the total left half space in front of the robot correctly as ”left” afterwards. An additional contribution is the mutual support between the learnt classes via the associations for improving the local decisions for learning and updating models. Other contributions are the integration of reflexive grasping for being able to condition an auditory label for command execution. This is being realized by the association mechanism as stated above. The system also features binaural far field audition for natural speech interaction without a close talk microphone that works during both ego and interactor motion, but this has already been reported in [8].

ALIS 3 is the latest system within our series. It is based on the architecture of ALIS 2 [7] and relies on following elements, as can also be seen in the figures 1 and 2: There are temporally stabilized basic visual, auditory, and tactile perceptions of the world, as well as a self collision free online motion generation and control system for the humanoid with an arbitration between several reactive behaviors. Those sensing and acting capabilities together provide favorable interactions between the robot and its surrounding world and form the basis for the learning and associations. The learning and the evaluation of the acquired concepts is uniformly governed by an expectation generation based behavior control mechanism. All learning is online and happens during interaction with the robot. The underlying capabilities and the architectural concepts have been published in [7]–[10].

With the presented system human tutors can freely interact without a close talk microphone. A tutor can teach interactively visual classes of relative positions to the robot like ”left”, ”bottom”, ”near”, he can teach different object size like ”large” and ”small”, and some other classes for the motion status of a proto-object or the height and orientation of planar surfaces like tables. The tutor can freely teach corresponding auditory classes, i.e. labels, and create associations linking the auditory labels to visually defined classes or basic behaviors for building internal concepts. The classes and labels can be trained independently from each other and at different times. The learned concepts can immediately be evaluated or used as commands. All learning and interactions are performed online during robot motions. To our knowledge this is the first time such kind of performance has been achieved with a full size biped humanoid.

The remainder of the paper is organized as follows: The next sessions will focus on the global architecture and the major elements. This is followed by a description of the learning of new classes and associations as well as the evaluation of the learnt knowledge within the interaction. The subseeding section is concerned with the experiments performed in interaction showing the researched learning capabilities. The paper ends with a short discussion and summary of the contribution.

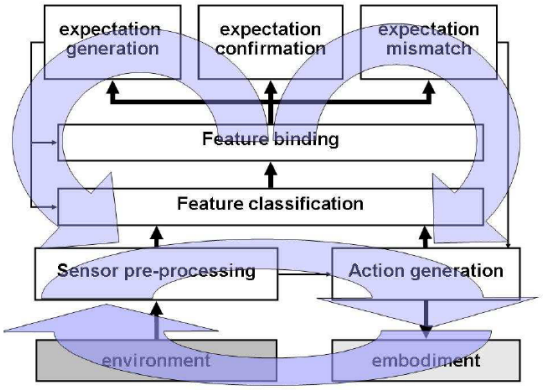


Fig. 1. Abstract architecture graph of ALIS 3

II. ARCHITECTURE AND ELEMENTS

The global architecture of ALIS 3 is sketched in figure 1. It can be subdivided into two main parts. The lower part is indicated by the two arrows forming a circle. This part is called the reactive layer. The remaining upper part is called abstraction layer and comprises the feature classification, the feature binding, and the expectation handling.

The reactive layer performs all basic sensory processing and basic behavior generation. It keeps the robot balanced, performs task level control, and establishes favorable interaction situations for the system by approaching and attending to selected Proto-objects. These Proto-objects are behaviorally relevant entities, obtained from a continuous decomposition of the world into basic sensory percepts without using any information about the identity of the underlying object [10]. Feature vectors, denoted by $\boldsymbol{f}_i^t$, represent a feature for the domain i at time t and serve as interface to the abstraction layer. A domain i is a currently prescribed basic perception, the ones considered in this work are: A combination of RASTA-PLP and HIST features for speech recognition ($i = 1$) [8], the visually determined 3D position ($i = 2$), the approximated size ($i = 3$), the motion status ($i = 4$) and some planar surface properties ($i = 5$) of a fixated proto-object as well as the activations of the basic behaviors as described below ($i = 6$). The system is equipped with several basic behaviors which reactively close the loop from the basic sensory processing back to the external world. The behaviors are arbitrated and can internally be biased for execution without the necessity of a corresponding external stimulus. The basic behaviors are: approaching, fixating and pointing at a selected proto-object, returning to the home position, grasping an object, releasing a grasped object as well as nodding and shaking the head. The behaviors have been described in detail in [7] except for the newly integrated grasping and releasing. The grasping behavior is based on the work presented in [11] and

is integrated in the following fashion. The hands of ASIMO are equipped with proximity sensors on the inside of the palm and on the outside. If the inside sensor is triggered by an object close to the palm the fingers are reactively closed. This represents a reflexive grasp like in little babies. If the outside of the hand is approached the fingers are opened. In order to start investigating physical manipulation within a comprehensive architecture, we also include a more powerful mode for grasping which is capable of planning and executing whole body motions for approaching and picking up previously defined objects. Nevertheless it must be activated either by the tactile sensors or by the behavioral bias.

The current activation of the behaviors forms the feature vector for $i = 6$. All feature vectors and the bias form the interface between the reactive and the abstraction part. The basic behavior generation as well as the generation of the feature vectors is performed reactively and continuously without the need of any cognitive control.

The abstraction layer derives and learns classes from the raw feature vectors and forms internal concepts by learning associations between the established classes. It generates and evaluates expectations about the perceived world by means of the associations. The match or mismatch of the expectations on the perception as well as on the actions side determine the external control flow wrt. actions and interactions. The novel internals of this layer will be described in detail in the next section.

The architecture has some important properties. The chosen feature representation provides a uniform data representation for all modalities (vision, speech, proprioception) and domains within the modalities. This is a crucial prerequisite for learning associations between and building concepts comprising arbitrary domains. The hierarchical subdivision into a reactive and an abstraction layer facilitates the learning from the viewpoint of complexity. The system is always responsive on a fast time-scale without cognitive control or planning and provides continuously the necessary features, representations and controls.

III. LEARNING, ASSOCIATIONS AND EXPECTATIONS

The internals of the abstraction layer and the interfaces to the reactive layer are depicted in figure 2. The abstraction layer performs the local learning of classes x in the visual and the speech domain based on novelty and a possible bias. Behaviors are currently not learned but associated. This layer also performs the learning of associations between class x from domain i and class y from domain j for all domains based on co-occurring observations. Which association is to be learned has to be specified by a learning mask, which currently has to be specified by the interactor. This eases the solution of the research question which association to learn in the case of several concurrently possible ones for now. Based on the learned classes and associations the system generates expectations and evaluates the match or mismatch for controlling the behavior.

The processing within the different domains is homogeneous except for some deviations in the behavior domain. The feature vectors are mapped to the classes yielding a memory activation $m^t_{i,x}(T_i)$ of the class x of domain i based on observations for the time span T_i. How the activations are determined depends on the domain. For speech they are derived from compounds of Hidden Markov Models, for vision they are derived from Gaussian models over the feature space, and for the behaviors they directly correspond to the current activations of the basic behaviors.

The activations have in common that they are based on local models. This is beneficial for representing observed data, but may be limiting for creating hypotheses about novel data i.e. for generalization. Therefore we distinguish between memory of recent observations and hypothesized assignments $h_{i,x}$ of features to classes. Those hypothesis activations are crucial for building broad concepts generalizing well beyond experienced observations. They are obtained by complete tessellation of the feature space to all classes in one domain:

$$h_{i,x} = \begin{cases} 1 & \text{if } x = \underset{y \in D_i}{\operatorname{argmax}}(m^t_{i,y}(T_i)) \\ 0 & \text{otherwise} \end{cases} . \qquad (1)$$

The effect is the following: imagine the system has learned the association between the speech label "near" and positions 60cm away from the robot and the speech label "far" and positions 120cm away from the robot. Without the hypotheses only positions sufficiently close to the original positions would be classified as "near" or "far", but not positions closer or further away from the originally learned ones. Based on the tessellation the hypotheses perform this kind of generalization. Diversifying and correcting wrong overgeneralizations is always possible during interaction. The hypothesis activations are used for evaluation purposes, the memory activations are used for novelty detection during learning. Both activations together are called hybrid data representation. The hypothesis activation is actually more advanced than described. For one domain several orthogonal groups can exist that can concurrently be activated. This means that an object can at the same time be "left" and "far". But further details are beyond the scope of this paper. The associations between the different classes are represented as 4-tuples $(i\,x, j\,y)$.

A. Learning of Classes

We will now describe the learning of classes. It is intertwined with the concept of a Learning Session. The session is indicated by the tutor. It serves two purposes: First, it provides an attentional mask a_i depicting the domains of interest in a way that $a_i > 0$ for attended domains and $a_i \leq 0$ otherwise. Second, the session specifies a time-frame $[t_0, t_1]$ in which feature vectors are assumed to belong to one distinct class per domain only. For example, one Learning Session can contain position-vectors for the left position and speech-features for the label "left", but it may not contain a combination of left and right position vectors. In our unsupervised learning approach, an essential part consists in deciding whether a

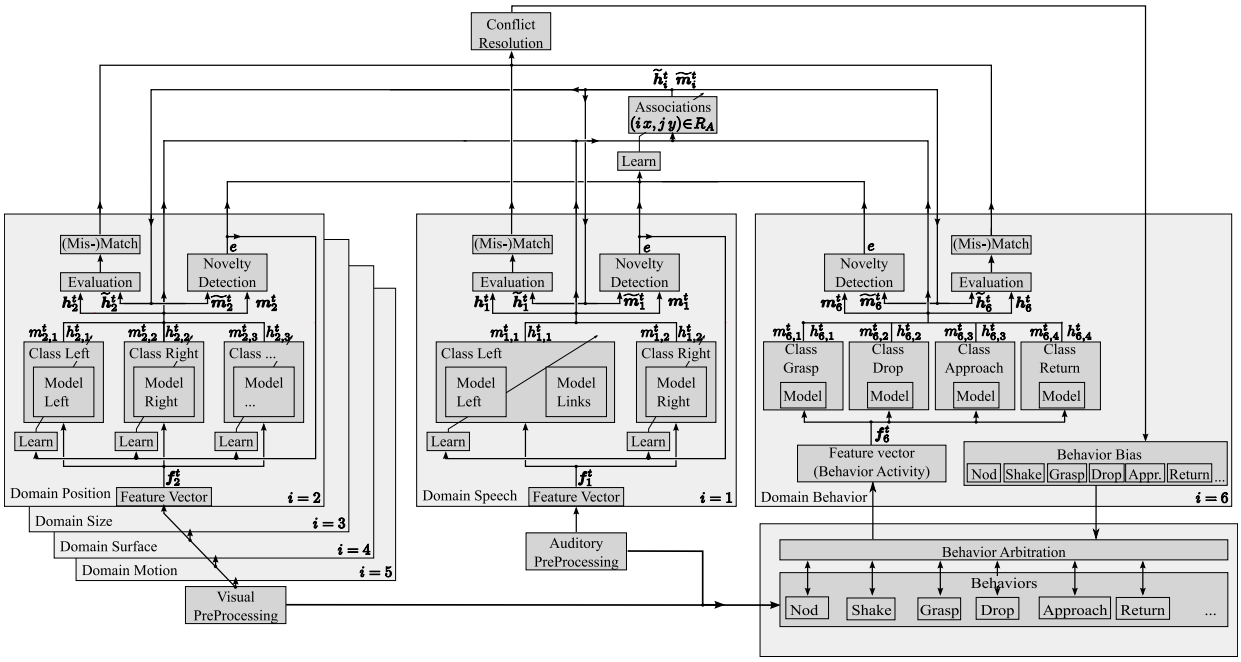


Fig. 2. Schematics of the abstraction layer including some elements from the reactive layer for exemplifying the interfaces.

presented stimuli belongs to a known class and should be used to update the underlying model, or whether it is unknown and should therefore be encapsulated in the model of a new class. We call this process Novelty Detection, it is based on two principles derived from biology. The first suggests basing the decision about new or known stimuli on the activation of the already learnt classes. A high class activation reflects sufficient coverage of the current stimuli by the existing classes - the residual is low, no new classes should be created. A low activation is an indication for insufficient representation, the residual is high and new classes are necessary. The Hypothesis Activations are not applicable here, because the tessellation of the feature-space abstracts from similarity to the underlying models. Therefore, this decision is based on the Memory Activations $m^t_{i,x}(T_i)$. For higher reliability, these activations are averaged over the complete Learning Session, so $T_i = t_0 - t_1$, where t_0 is the time the learning-session started and t_1 is the time it stopped.

The second principle suggests that we do not base the Novelty Detection on the activations in domain i only, but rather incorporate the activation of all associated classes by introducing the top-down activation $\widetilde{m}^t_{i,x}$:

$$\widetilde{m}^t_{i,x} = \sum_{(j,y)|(i\,x,j\,y)\in R_A} w_j \cdot m^t_{j,y}(T_j) \ , \qquad (2)$$

where $w_j \in [0,1]$ is a weighting to steer the influence of domain j. The weights are a-priori estimations of the stability of the features from the different domains. They can be derived individually for the different domains.

If the collected feature-vectors belong to a known class they should be well represented, indicated by a significant top-down activation $\widetilde{m}^t_{i,x} > \Theta_K$. In contrast, features of an unknown class will rarely be associated to an existing model and therefore result in a top-down activation $\widetilde{m}^t_{i,x} \leq \Theta_K$. For this binary classification between new or known classes, an evaluation of all classes in one domain is unnecessary. The investigation of the class l with the highest top-down activation is sufficient. The result of this Novelty Detection is a teaching signal e that either depicts the class to be adapted or indicates the creation of a new class:

$$e = \begin{cases} \underset{x\in D_i}{\operatorname{argmax}}\, \widetilde{m}^t_{i,x} & \text{if } \max\limits_{x\in D_i}(\widetilde{m}^t_{i,x}) > \Theta_K \\ N_i + 1 & \text{otherwise} \end{cases} .$$

The number of currently existing classes for domain is denoted with N_i. The adaptation of existing classes depends on the implementation of the class representation and is not covered here. The described learning mechanisms are conceptually homogeneous for all domains, as can be also seen in figure 2. This is a major step towards general cross modal learning.

B. Learning of Associations

We will now focus on the learning of associations based on co-occurring observations. We prefer the term "co-occurring" to "synchronous" because the different domains may operate in different time scale without a strict synchronization. In principle, there are many different ways in finding these associations, but we require the system to learn from very few examples. This prevents the use of purely statistical methods, like e. g. Hebbian or correlation learning. Detecting correlations between the Memory Activations during a Learning Session is possible and would also work with few examples, but feature vectors belonging to yet unrepresented classes will show no activations and can therefore not be correlated.

To avoid these problems and master the mentioned requirements, we utilize the result of the Novelty Detection, depicting the index of the best matching class or a new class. This index exists in two different domains, the speech-domain and another domain depicted by the attentional mask a_i. To support the learning between more then two domains, we collect the tuples (i, e) for all unmasked domains i with $a_i > 0$ and store them in a set L. A combination of all tuples in L is then added to the set R_A to represent the connections between these classes:

$$R_A = R_A \bigcup_{\substack{(i,x) \in L \\ (j,y) \in L}} (i\,x, j\,y) \tag{3}$$

The presented method for learning associations can cope with several different kinds of tuples L based on the abstracted result of the Novelty Detection. There can be new associations between know classes, new associations between know and new classes and totally new classes and associations.

C. Evaluation and Expectations

After describing how the learning works we will now look at the evaluation of the learnt classes and concepts. The core mechanisms governing those internal processes are called expectation generation, match evaluation, and mismatch resolution. Similar to the learning they are conceptually homogeneous for all domains.

The core mechanism is based on the comparison of features from different domains. This is for example necessary to let the system decide if an understood speech-label matches any of the perceived visual classes. The explanation given here is based on our work presented in [9]. A comparison between the different domains at the level of feature-vectors is impossible due to the qualitative difference, or more precisely: the feature-space, dimension, and time-representation. For this reason the system utilizes the introduced Hypothesis Activations to compare whether for example the active speech-class matches one or more visual classes. In a first step the Hypothesis Activations $h_{i,x}$ are computed for each domain i using the current feature-vector $\boldsymbol{f}_i$ according to (1). In the next step, the activity of all classes in one domain must be compared with the activity of all associated classes. This comparison should result in a "match" if for an active class x in domain i the associated class y in domain j is also active. In contrast, it should result in a mismatch, if for an active class x in domain i the associated class y in domain j is inactive and additionally, there exists an active class l in domain j which is associated to a class k in domain i, with $k \neq x$.

Let us assume, for example, there are two position classes, one representing the left position, and the other one representing the right position. Let us further assume, that the position-class left is associated to the speech-class left and the position-class right is associated to the speech-class right. If position-class left is active the comparison results in a match, if the speech-class left is active, too. It results in a mismatch, if in contrast the speech-class right is active.

Mathematically this comparison can be formalized as a similarity measure between two activation vectors. The first vector $\boldsymbol{h}_i$ is a concatenation of Hypothesis Activations in domain i:

$$\boldsymbol{h}_i = (h_{i,1}, \ldots, h_{i,N_i})^T \ . \tag{4}$$

The second vector is the top-down hypothesis vector $\widetilde{\boldsymbol{h}_i}$:

$$\widetilde{\boldsymbol{h}_i} = (\widetilde{h}_{i,1}, \ldots, \widetilde{h}_{i,N_i})^T \ , \tag{5}$$

with

$$\widetilde{h}_{i,x} = \sum_{\substack{(j,y) | (i\,x,\, j\,y) \in R_A \\ \wedge\ (j,y) \neq (i,x)}} h_{j,y} \ . \tag{6}$$

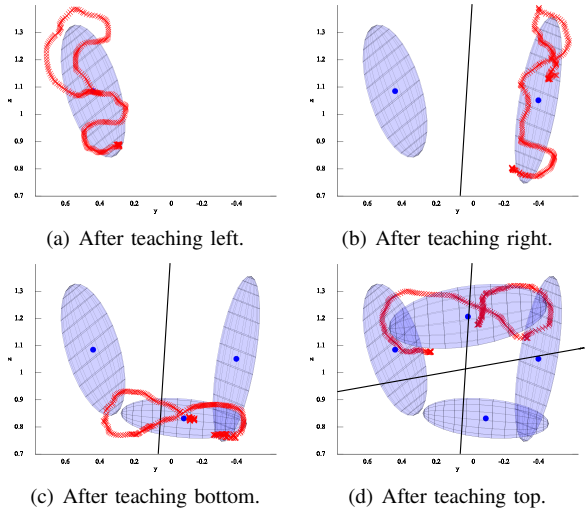
The top-down Hypothesis Activation $\widetilde{h}_{i,x}$ is a summation of all Hypothesis Activations associated with $h_{i,x}$ via a 4-tuple in R_A, but excluding $h_{i,x}$ itself. In a match situation, the top-down hypothesis vector can thus have a different length, but point in the same direction as the hypothesis vector $\boldsymbol{h}_i$. Hence, a suited measure of similarity is e. g. the scalar product $d(\boldsymbol{h}_i, \widetilde{\boldsymbol{h}_i})_S$ between $\boldsymbol{h}_i$ and $\widetilde{\boldsymbol{h}_i}$. If the similarity between these two vectors falls below a significance threshold Θ_C, the system treats this as a mismatch in domain i, and if the similarity raises above Θ_C, this is interpreted as match in domain i.

The match / mismatch computation is local to the domains. The top-down influence is currently directly computed via the association. For the current state of the system this is sufficient, because the focus is here on the learning of classes and the building of concepts. But the architecture is already prepared for receiving top-down influences form other sources than the associations. Those top-down influences then represent higher cognitive expectations or goals. Such a next level of the architecture that autonomously controls the creation of such expectations or goals is subject to current research.

Currently the set of all match evaluations can directly be used for evaluating the learnt classes and concepts or employing them as commands. Here we have to distinguish between mismatches in the perceptive domains and the behavior domain. For both the mechanisms are rather straight forward.

A perceptual match or mismatch occurs if the tutor presents the system something with a learnt visual property class and utters a label. If they match a bias for the nod behavior for ASIMO's head is generated and the evaluation is finished. If they don't match a bias for the shake behavior for ASIMO's head is generated and the reactive layer is forced to attend to something different while keeping the expectation raised by the label. This kind of behavior is called conflict resolution. Please note that in principle any kind of conflict resolution could be triggered including the active search for a specific object with the verbally specified visual property. For now the systems just communicates the state of the match by means of gestures for "yes" (nod) and "no" (shake) and relies on the interaction in order to experience a match.

A behavioral match or mismatch occurs if the tutor utters a label that is associated with a specific behavior. If the



(a) After teaching left. (b) After teaching right.

(c) After teaching bottom. (d) After teaching top.

Fig. 3. Models of learned positions left, right, bottom and top. Please refer to text for details.

associated behavior is already active it is match and the evaluation finishes. If it is a mismatch a bias for the corresponding behavior is generated by the conflict resolution and communicated to the behavior arbitration for activating the corresponding behavior controller.

A deeper understanding of the described learning and control mechanisms should be gained from the following description of the interactive experiments.

IV. Experiments

In this chapter we will demonstrate the feasibility of the presented system in two representative experiments. These experiments are documented in an accompanying video.

The experiments focus on the learning capabilities of the system, putting special emphasize on the visual learning and the overall system performance. In other experiments we evaluated the stability of the visual perception and the speech learning capabilities [10], the robustness of the far field audition [8], and the coupling of reactive and cognitive layer [9].

In the first experiment, we will focus on the visual learning of position classes, namely left, right, bottom and top. This experiment will demonstrate the autonomous learning of classes and handling of orthogonal groups. The second experiment gives a global system view by monitoring the system states during the learning of audio-visual concept for small and large, as well as audio-behavior concepts for "take" and "release". It also demonstrates the ability to learn independently for each modality based on the individual Novelty Detection. Both experiments are carried out using a Honda ASIMO robot.

Description Experiment 1: In the first experiment a tutor steps in front of the robot, utters a predefined speech command to start the learning session for position learning and teaches the robot the position left by presenting an object in the left visual viewfield of the robot. To cover a representative area of valid stimuli for left, the tutor moves the object around. In Figure 3(a) the red crosses indicate the presented positions in the robots heel coordinates, with the x-axis pointing forward, y-axis pointing left, and z-axis pointing upwards. After this learning session, the tutor moves to the right side of the robot and repeats the steps described above for the learning of "right". Again the presented stimuli are visualized as red crosses in Figure 3(b). The Figures 3(c) and 3(d) contain the learning samples for the learning of bottom and top.

Results Experiment 1: At the end of each learning session, the system creates a local model to approximate the presented stimuli by approximating mean and covariance. These models are visualized in the Figures 3(a)-3(d) after each learning session, where the blue dot presents the mean and the blue ellipsoid the covariance. As described in section III the system estimates hypothesized assignments based on the local models, leading to a tessellation of the feature space. The discrimination border is visualized as a black line in the Figures above. Estimating a discrimination border requires at least two models, therefore it is not visible in Figure 3(a). However, learning "bottom" does obviously not affect the discrimination border for left and right. In contrast, an additional classification border is visible in Figure 3(d). This independent treatment of left and right from bottom and top shows the systems ability to detect the independence of horizontal- from vertical positions.

Description Experiment 2: A visualization of the tutor's and robot's actions in this experiment is given by the images in the top-row of Figure 4. To demonstrate the systems ability to learn independently for each modality, the tutor starts teaching the *visual* class for small by presenting a small cup but without saying anything($t = 1$). Afterwards, he teaches ASIMO the *word* "small" but without showing him something small($t = 2$). In the following step he lets the system learn the association between the two, by presenting something small and uttering the word small simultaneously($t = 3$). Here, the Novelty Detection should classify the presented stimuli as known and adapt the existing classes for small.

In the next step the tutor teaches the system the concept of "large" by uttering the word while presenting a jar ($t = 4$). Afterwards he evaluates the learned concepts by presenting the jar and saying "large", as well as presenting the cup and saying "small"($t = 5$). In the next step he examines the systems ability to generalize by placing a table in front of the robot and evaluating it as "large" ($t = 6$). Subsequently the reflexive release and grasp is demonstrated by touching the proximity sensors at the outside of the hand($t = 7$) and on the inside($t = 8$). To teach the coupling between the word "release" and the releasing action, the tutor starts the action-learning, activates the outside proximity sensor to trigger the release action and utters the word "release" several times ($t = 9$). He then repeats the process for the word "take" by activating the inside proximity sensor ($t = 10$). Finally he uses the thus created associations to let Asimo take a basket ($t = 11$) and release it($t = 12$).

Results Experiment 2: Row (2) of the plot in Figure 4

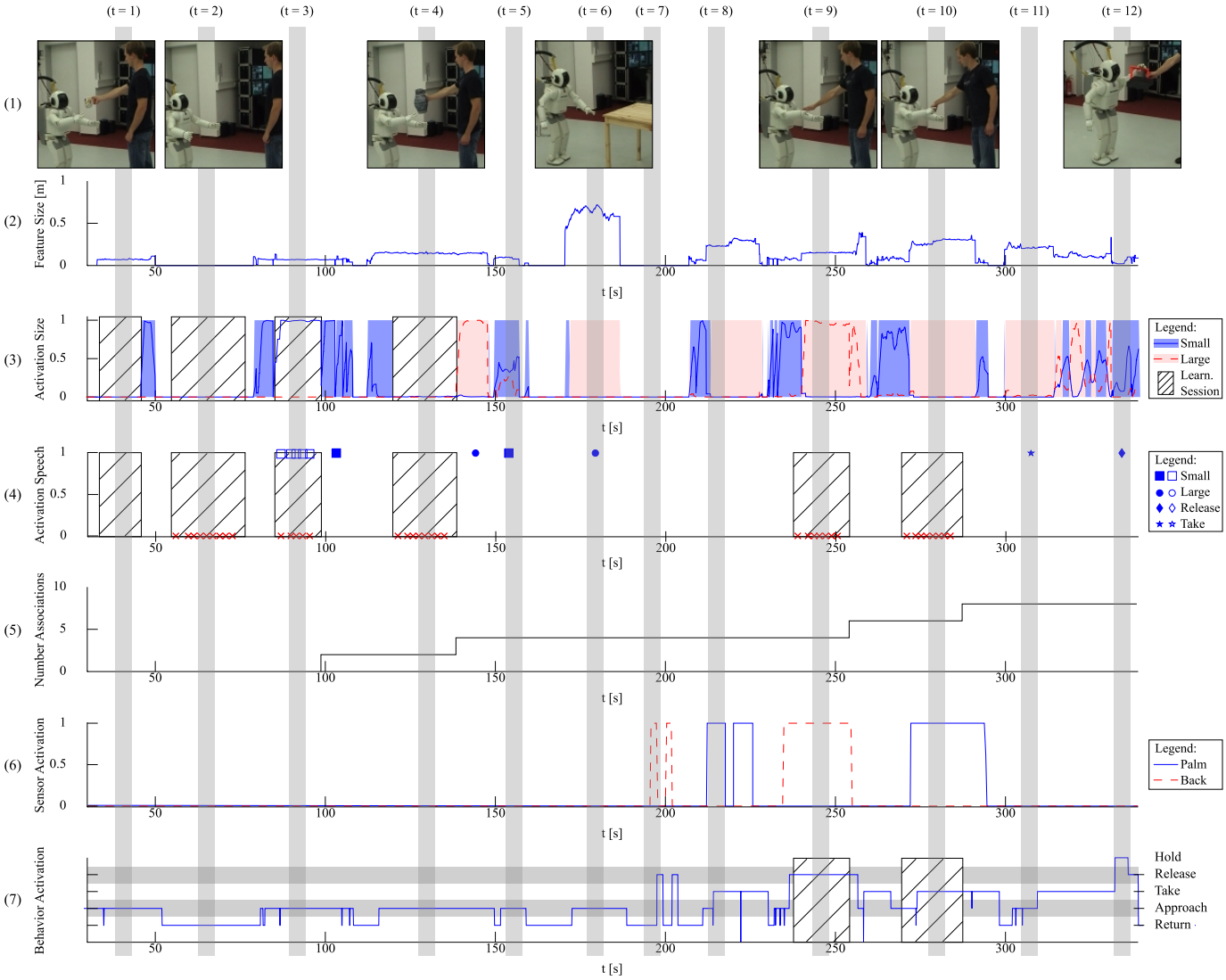


Fig. 4. Plot of system states. Please refer to text for details.

shows the measured 3D-size of the presented object. A zero-line indicates phases of the interaction, where no object is presented to the robot. Row (3) contains the activation of the visual size models, the lines represent the memory activations and the filled areas the hypothesis activations. The activations of the model small are visualized in red, the ones for large in blue. Additionally, this plot visualizes the learning session as white striped area.

At time $(t = 1)$ the learning session is active, but no model is learned yet. From row (2) we can infer the presentation of a ca 8cm big object. Row (4), visualizing the activation of the speech classes, shows the active learning session, but otherwise no speech activity, because the tutor is not speaking. Consequently after this learning only a new visual class is created, indicated by the raising size-activity. At time $(t = 2)$ row (4) contains a lot of unrecognized speech activity (visualized as red cross), but no object is presented and therefore no size-activity is visible in row (3). During the learning session at time $(t = 3)$ the visual class for small is active as well as the speech class for “small”, with the memory activation of the latter visualized as square in row (4). Please note, that after this session, only the number of associations changes from 0 to 2 (two, because associations are symmetrical), but neither new visual classes nor new speech classes emerge.

The learning of “large“ using a ca 14cm big jar leads to the creation of a new visual- and a new speech class and their association. During the evaluation of ”large“ and ”small“$(t = 5)$, the recognized speech classes match the visually perceived sizes. While the tutor evaluates the table as ”large“ at $(t = 6)$, the memory activations for small and large in row (3) drop close to zero, caused by the relatively large size of the table (ca. 68cm) compared to the previously taught objects. However, the hypothesis activation of the class large remains active.

Row (8) contains the activations of the robot behavior. Note the correlation of the behavior ”return“ in the latter row with

the 0-line of the object size: If no object is presented, the robot returns to his starting position. In contrast, whenever he perceives an object, he approaches it. In row (6) the activation of the palm proximity sensor is visualized as a blue line and the outside sensor as red line. At time($t = 7$) the reactive triggering of the "release" behavior using the outside proximity sensor is visible. The same holds for the reactive grasping triggered by the inside sensor at time ($t = 8$). During the learning of the words "release"($t = 9$) and "take"($t = 10$) the learning session is again visualized as striped area in row (7) and row(5). Observe the raising number of associations after each learning step. The effect of uttering the words "take" ($t = 11$) and "release" is visible in the plot: The respective behaviors are activated from the recognized speech commands, without the presence of any sensor signals.

V. Discussion

Together with functional evaluations in previous work [7]–[10], the results presented in the previous section close the loop of evaluating our system. We demonstrated the feasibility of learning visual models to approximate experienced perception, and showed the reliability of these models to classify stimuli as new or known.

As demonstrated in Experiment 2, the combination of local and global class-representations enables the system to generalize learned concepts to completely new situations. Obviously this might lead to overgeneralization whereas relying purely on observed data appears relatively safe. However, we prefer using the generalizing global representations, because they meet our understanding of intelligence as the application of known concepts to new situations.

We also presented our systems ability to learn synchronously in multiple domains, e.g. learning new words and new visual models. In this sense our system is superior to state-of-the-art systems as the one presented by Roy et al. [4]. In contrast to Iwahashi [5] we showed that our system does not *require* this kind of synchronous presentation, because it can learn individually in all domains and associate concepts, once it observed them together.

Finally, we demonstrated that the learning of concepts is not limited to audio-visual associations but also reliably works for the learning of actions. This multimodal integration of vision, audition, and actions is the prerequisite for the learning of behavior oriented representations.

One drawback of the presented learning mechanism lies in the fixed associations, which can not be changed once they are learned. Changing this would require continuous reevaluation of associations based on statistical observations, which would lead to a much higher number of training samples.

From a developmental point of view, there are two major aspects of our system design, which are pre-determined and not learned by the system: Triggering a learning session using a predefined speech command and class representations of the robots behavior. Both aspects are focus of our current research.

VI. Summary

In this paper we presented the current instance of our ALIS architecture, which is addressing the long term goal of creating a developing cognitive system. We also presented our latest research results wrt. learning and associating learnt knowledge to concepts. The approach we have presented performs in real-time in interaction with a tutor. The learning can proceed sequentially or in parallel for new visual properties, auditory labels and associations between the acquired classes. The newly introduced hybrid representations of the classes yield a broad grounding and permit stable mutual support between the classes for global decisions. Future work will cover the learning of behaviors as well as a more cognitively oriented control of the evaluation of the acquired concepts including more manipulation skills.

VII. Acknowledgments

The authors would like to thank Ursula Körner, Marcus Stein, Antonello Ceravola, Martin Heracles, Mark Dunn, Frank Joublin and Edgar Körner for their contributions, support and advice.

References

[1] D. Vernon, G. Metta, and G. Sandini, "The icub cognitive architecture: Interactive development in a humanoid robot," in *Proceedings of the IEEE 6th International Conference on Development and Learning (ICDL)*. Monterey, California, USA: IEEE Press, 2007.

[2] V. Krunic, G. Salvi, A. Bernardino, L. Montesano, and J. Santos-Victor, "Affordance based word-to-meaning association," in *IEEE International Conference on Robotics and Automation (ICRA 2009)*. IEEE, 2007.

[3] H. Jacobsson, N. Hawes, G.-J. Kruijff, and J. Wyatt, "Crossmodal content binding in information-processing architectures," in *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, The Netherlands, March 12–15 2008.

[4] D. Roy, *A mechanistic model of three facets of meaning*. Oxford University Press, 2008.

[5] N. Iwahashi, "Robots that learn language: Developmental approach to human-machine conversations," in *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication*, P. Vogt and et al., Eds. Springer, 2006, pp. 143–167. [Online]. Available: http://www.isrl.uiuc.edu/ amag/langev/paper/iwahashi06robotsEELC.html

[6] C. Goerick, "Towards cognitive robotics," in *Creating Brain-like Intelligence*, ser. LNCS 5436, B. Sendhoff, E. Koerner, O. Sporns, H. Ritter, and K. Doya, Eds. Springer Verlag, 2008.

[7] B. Bolder, H. Brandl, M. Heracles, H. Janssen, I. Mikhailova, J. Schmüdderich, and C. Goerick, "Expectation-driven autonomous learning and interaction system," in *IEEE-RAS International Conference on Humanoids*, 2008.

[8] M. Heckmann, H. Brandl, J. Schmuedderich, X. Domont, B. Bolder, I. Mikhailova, H. Janssen, M. Gienger, A. Bendig, T. Rodemann, M. Dunn, F. Joublin, and C. Goerick, "Teaching a humanoid robot: Headset-free speech interaction for audio-visual association learning," in *Proc. 18th IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*. Toyama, Japan: IEEE, 2009.

[9] I. Mikhailova, M. Heracles, B. Bolder, H. Janssen, H. Brandl, J. Schmüdderich, and C. Goerick, "Coupling of mental concepts to a reactive layer: incremental approach in system design," in *Proceedings of the 8th International Workshop on Epigenetic Robotics, Brighton, UK*, 2008.

[10] J. Schmüdderich, H. Brandl, B. Bolder, M. Heracles, H. Janssen, I. Mikhailova, and C. Goerick, "Organizing multimodal perception for autonomous learning and interactive systems," in *IEEE-RAS International Conference on Humanoid Robots*, 2008.

[11] M. Gienger, M. Toussaint, N. Jetchev, A. Bendig, and C. Goerick, "Optimization of fluent approach and grasp motions," in *International Conference on Humanoid Robots*. IEEE Press, 2008.