7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Multimodal feedback in auditory-based active scene exploration

Thomas Walther
Institut für Kommunikationsakustik, Ruhr-Universität Bochum, Germany.

Benjamin Cohen-Lhyver
Sorbonne Universités, UPMC Univ. Paris 06, UMR 7222, ISIR, F-75005 Paris, France
CNRS, UMR 7222, ISIR, F-75005 Paris, France

Summary
Human cognition still easily walks over existing technical solutions when it comes to audio-visual scene analysis. In contrast to most technical approaches, the biological paradigm employs sophisticated bottom-up/top-down processing of the observed input, making intense use of cortical feedback loops. Within the Two!Ears project framework, we propose to artificially mimic such biological feedback from an 'active' listening perspective. In the current contribution, we deal with ongoing work in the context of enabling robots to assist in the exploration of search-and-rescue scenarios, where the machines augment auditory localization cues with visual information to precisely fix the positions of human victims. Our experimental robotic platform is enabled to perform translatory movements in the horizontal plane and carries a replica of a human head, which can tip and rotate and is equipped with 'eyes' and 'ears'. The sensor signals are handled by a system that incorporates both bottom-up and top-down processing of the observed input, making intense use of cognition-based feedback loops. Going well beyond familiar feed-forward localization algorithms, feedback loops between the audio-visual sensor arrays and the robot's motor-control unit allow for 'active' exploration of given scenarios. In cases where incoming acoustic target signals cannot directly be observed, for example, signals from a victim hidden behind debris or walls, the robotic platform adopts its motion patterns very much like a human being and visually peeks around the obstacles to verify or falsify initial hypotheses. Currently, our work is focused on setting-up the feedback and communication architecture for the necessary complex binaural and visual evaluation system, comprising signal-processing as well as symbol-processing stages.

PACS no. xx.xx.Nn, xx.xx.Nn

1 Introduction

Following the insight that '*exploration and search* is a typical task for autonomous robots performing in rescue missions' [1], many approaches for solutions have evolved in this field over the past years. However, the vast majority of these approaches addresses the underlying problem from a purely technical perspective, neglecting potential benefits that could be gained by learning from biology. The Two!Ears project amends the latter issue by introducing artificial cognitive feedback into a standard search and rescue (S&R) system, enabling a robotic front-end to mimic human behavior in complex scenarios, based on the extended use of multimodal information. The following section surveys current trends in robotic S&R applications, particularly those which focus on active exploration and multimodal aspects. Beyond that, we shall guide the readers' attention to simulated biological feedback and the presence of *organic computing* [2] (OC) strategies in contemporary robotic systems.

1.1 Robotic rescue systems

When surveying the relevant literature, it comes clear that the idea of using robots in disaster scenarios has extensively been pursued over the last decades. To stay in scope, we focus on standard land-based robotic rescue systems in the following. Nevertheless, note that recent aerial, naval and combined solutions (e.g. [3, 4]) underline the potential of robots in perilous environments. With form following function, land-based S&R robots show a large variety in shape and size: snake-like, 'hyper-redundant' [5] machines are enabled to easily pass tubes or other small openings to find entrapped persons. Quadro-, hexa-, or octapod robots

(c) European Acoustics Association

can be employed in difficult terrain or urban emergency to find victims under debris (see, for instance, [6]). Whereas the aforementioned devices have a relatively restricted operation area, wheeled robots lend themselves for large-scale S&R tasks (see, e.g., [7]). Though such wheel-based platforms work reliably in streets or flat open country, they are easily stopped by wreckage or rocks, unless hybrid solutions are employed, see [8]. Allowing a number of small mobile devices to team up leads to a significant efficiency increase when it comes to the task of localizing potential victims [9]. Eventually, humanoid robots combine flexibility with range: mimicking human motion capabilities, these machines overcome complex obstacles, e.g., staircases [10], and show the potential to roam extended territories in long-range S&R tasks.

1.2 Exploration and cognitive feedback

Following the insight that 'exploration refers to the act of gathering information about an environment as result of being curious' [11], artificial curiosity is a long-known [12] and ongoing method to install active exploration capabilities in robots, be it for active environmental exploration, or autonomous learning of certain motion patterns [13, 14]. However, unbiased exploration of task spaces is likely to fall foul of weak performance. In fact, several authors note that exploration should be guided in such a way 'that the expected knowledge gain is maximal' [12]. Basic examples for such *directed* exploration approaches may be found in [12], or [15]. To enable active exploration in humanoid robots by combining 'social guidance and intrinsic motivation' [13], Ivaldi introduces a hierarchical model: first, *task space exploration* techniques decide on which goal to pursue, then *state space exploration* mechanisms try to approach this goal by appropriately actuating the robotic device. Inherent to this novel approach is the robot's ability to mimic human feedback strategies: multimodal sensor input is fed back into the planning system, allowing for 'the emergence of visuo-motor representations and cognitive loops' [13]. Staying in the feedback context, [16] coin the term *cognitive synergy* to describe the feedback loops in their generic robot control architecture. Other approaches propose combined use of bottom-up/top-down techniques to endow robots with 'child-like intelligence' [16]. [17] also follow such a bottom-up/top-down strategy in order to guide their mobile robot's visual focus. It has to be emphasized that the idea of evoking human-like behavior in machines by installing artificial cognition and appropriate feedback loops is closely related to *organic computing* ideas. OC approaches attempt to render artificial devices closer to organic creatures, using, inter alia, biological principles of *self-organization* and *generalization*. Lessons learned from the organic computing domain can hence become useful in the context of Two!Ears. For a more detailed overview of OC strategies, [2] lends itself.

1.3 Multimodality in robotics

As 'the richness of perceptual information is a hallmark of humans' [13], many authors employ multimodal approaches to supply robotic cognition with as many sensory impressions as possible. For such techniques, *cue fusion* becomes mandatory: herein, 'the key idea is that it is safer and more stable to decide based on multiple low thresholds of different nature that complement each other, than using a single highly tuned one' [18]. In many multimodal systems, a combination of color and depth information still dominates, see [19], [20] or [21]. Coming closer to ideas found in the Two!Ears project framework, [22] uses a biologically-inspired combination of sound and baseline vision information to enable object detection in a mobile robot. [23] exploits sound and texture information to locate humans in a constrained scenario. [13] guides learning in their humanoid robot by combining color, depth, sound and proprioceptive cues. As can be inferred from the latter approaches, multimodal cue fusion plays a leading role in contemporary robotics; we follow this intuition by endowing our system with the ability to operate on an extendable array of cues, including color, emulated depth, and (emulated) sound.

2 System overview

As the Two!Ears system, as proposed above, will evolve over a time span of almost three years, many elements required for feedback implementation and testing can not realistically be expected to be immediately available. This is particularly true for complex elements that deal with abstract functions, like object labeling and/or scene understanding. Also, the need for a physical robotic front-end is unlikely to be satisfied in the first stages of the Two!Ears project. To address these issues, the *Bochum Experimental Feedback Testbed* (BEFT) as introduced in the current paper represents a self-contained test environment that allows to probe a variety of feedback strategies early in the project's life-cycle. The following sections describe BEFT in a more detailed way.

2.1 Virtual environment

As announced above, the BEFT framework rests on virtual representations of baseline search and rescue scenarios $\mathcal{S} = \{\mathbf{s}_i\}, i = 0...N_S - 1$. Each of these scenarios is described via XML and consists of a list of environmental objects $\mathcal{O} = \{\mathbf{o}_i\}, i = 0...N_O - 1$, which carry sets of simulation-related parameters, i.e., $\mathcal{P}_i = \{\mathbf{p}_i, \phi_i, l_i, \mathbf{v}_i, \mathbf{a}_i\}$. Herein, $\mathbf{p}_i$ and ϕ_i describe the position and orientation of an object $\mathbf{o}_i$ in 3D space. Let $\mathcal{K} = \{k_i\}, i = 0...N_K - 1$ define a set of available

categories, with $k_i \in \{person, car, siren, wall\}$. With that, be $l_i \in \mathcal{K}$ the *category label* of $\mathbf{o}_i$. Further, $\mathbf{v}_i$ is $\mathbf{o}_i$'s *visual representation*, corresponding to an entity maintained in the *Ogre* 3D [24] render engine. *Ogre* provides a powerful framework for real-time visualization and allows for visual inspection of all scenario contents.

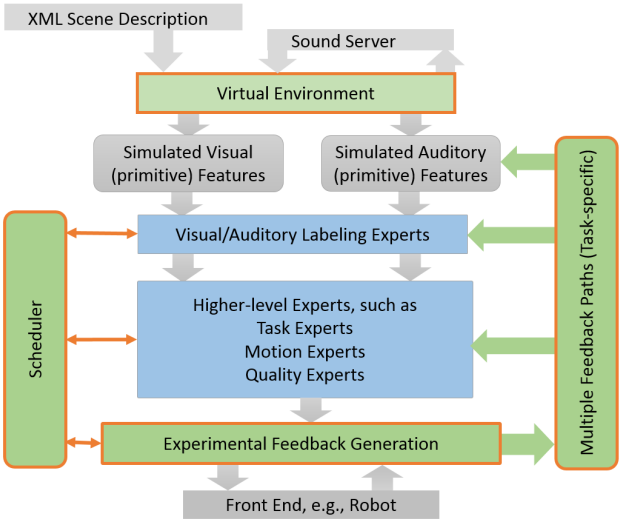

Figure 1: The *Bochum Experimental Feedback Testbed*

In addition, $\mathbf{a}_i$ conveys an *acoustic representation* for each object i. This acoustic representation provides a *sound stimulus* file tag t_i (basically a link pointing to a '.wav' sound file), and an *activity* flag $f_i \in \{0, 1\}$ which controls the acoustic behavior of an object i. $\mathbf{o}_i$ emits sound iff $f_i = 1$. Note that each virtual object i has a corresponding *acoustic object* $\hat{\mathbf{o}}_i$. This object maintains the estimated *acoustic position* $\hat{\mathbf{p}}_i$ and the inferred *auditory category label* $\hat{l}_i$ for each $\hat{\mathbf{o}}_i$. At this point of discussion, both $\hat{\mathbf{p}}_i$ and $\hat{l}_i$ are derived by plainly deducing from $\mathbf{p}_i$ and l_i. While this strategy is sufficient in the current scope, the BEFT environment maintains a TCP/IP interface that enables straightforward integration of external *auralization* software provided by our project partners, like the *Sound Scape Renderer* [25]. Combining such enhanced auralization mechanisms with adequate estimation strategies will allow for 'true' acoustic inference later in the course of the project.

2.2 Robotic front-end

To enable active exploration of virtual environments, the BEFT system embodies a movable *robot platform* object that mimics a physical PR2 [26] robot device. The original head of the PR2 is removed and replaced by a virtual representation of a standard *KEMAR* [27] artificial head, the human-like *HRTFs* [28] of which fulfill the specific requirements of the Two!Ears project. CAD data for the PR2 platform stem from [29], whereas – due to a lack of adequate CAD material – the KEMAR head was constructed using the *MakeHuman* plugin [30], which is available within the *Blender* [31] software framework. While the generated KEMAR model visually deviates from the original KEMAR head, we will employ the original KEMAR's HRTFs for auralization in later system generations. Thus, the visual difference can safely be neglected from the point of view of acoustics. To allow for multimodal scene analysis, the virtual KEMAR is equipped with an artificial RGBD camera (color+depth, realized via specific GPU functions).

2.3 Task assignment

Control of the robotic object is based on a dedicated *task stack* $\mathcal{T} = \{t_i\}, i = 0...N_T - 1$. The spectrum covered by the individual t_i's extends from basic motion commands to extensive high-level tasks, like, for instance, retrieval of all human victims in a given scenario. Thereby, tasks of increased complexity autonomously launch less complex sub-tasks on demand. Current task execution can be interrupted at any time by either pushing a task of higher priority onto the stack or by clearing the task memory. Note that tasks can be either injected manually into the processing order, or are issued by the expert system described below.

2.4 The hierarchy of experts

The virtual robotic front-end employs a hierarchical *system of experts* $\mathcal{E} = \{\mathbf{e}_i\}, i = 0...N_E - 1$ to evaluate the above parameters and to 'assign meaning' (compare [32]) to the observed objects. To that end, a dedicated expert $\mathbf{e}_i$ can take on different roles: *low-level experts* operate on the level of *primitive features* and read simulated physical data (e.g., $\mathbf{p}_i$, $\hat{\mathbf{p}}_i$) while generating output information such as $\mathbf{o}_i$'s or $\hat{\mathbf{o}}_i$'s distance/acoustic azimuth relative to the virtual robot. This information is forwarded to *mid-level experts* that work on the *symbol-processing layers* and try to assign meaning to each encountered $\mathbf{o}_i, \hat{\mathbf{o}}_i$, based on $l_i, \hat{l}_i$ and specific *sensor degradation functions* which will be described below. Eventually, *high-level experts* receive information from low-level and mid-level processing stages in order to perform sensor fusion, that is, integration of visual and acoustic cues. In this way, feedback loops will be established to simulate complex robotic behavior, such as active exploration. It has to be emphasized at this point, that our low- and mid-level experts are intrinsic to the overall structure of the BEFT, whereas the higher-level experts can be supported and/or replaced by external expert systems on demand.

2.5 Scheduling

To orchestrate the above ensemble of experts, we employ *Petri net* [33] scheduling. This practice is common in contemporary robotic systems (see, e.g. [1])

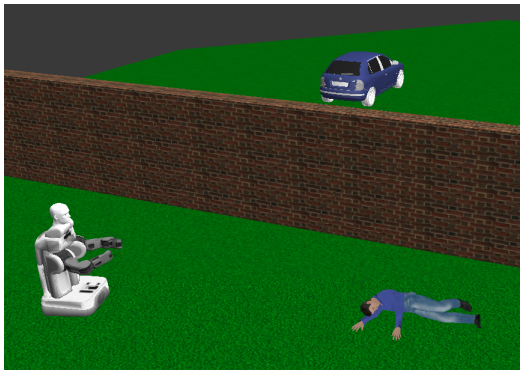

Figure 2: A typical BEFT search & rescue scenario

and allows for setting up schedules of nearly arbitrary complexity, ranging from standard sequential patterns to complex parallel strategies. An expert $\mathbf{e}_i$ is activated iff the corresponding transition $t_{\mathbf{e}_i}$ in the task-related Petri net 'fires' [33]. Currently, our expert schedule fires low-level experts simultaneously, followed by a conjunct activation of all mid-level experts. Then, the high level experts are enabled in a sequential manner. This strategy respects the hierarchical relations of all $\mathbf{e}_i \in \mathcal{E}$ and can easily be extended for more complex scenarios.

2.6 Degradation functions

The above experts could derive the parameters for each $\mathbf{o}_i$ with absolute precision using ground-truth information stored in $\mathbf{s}_j$ and could easily forward this information to the virtual robotic front-end. In contrast, the physical PR2 instance is going to face real-world issues, such as limited camera resolution, visual/acoustic occlusion of objects, concurring sound sources, and has only very limited access to a priori scenario information. As we aim at testing our feedback loops in realistic conditions, we thus have to artificially degrade available synthetic information. To that end, the low- and mid-level experts incorporate specific *degradation functions* to retrieve an object's *visibility* and *audibility*. Assume that $\mathbf{d}_i$ is a vector extending from the PR2 platform's center to some virtual object i. Let $d_i = \|\mathbf{d}_i\|$ be the corresponding distance between the robot and $\mathbf{o}_i$. Further, be $\alpha_i \in [-180°; 180°]$ the azimuth angle between $\mathbf{d}_i$ and the 'nose tip'-unit vector $\mathbf{n}$ of the simulated KEMAR head. Note that α_i is 0° iff the scalar product between $\mathbf{u}_i = \mathbf{d}_i/\|\mathbf{d}_i\|$ and $\mathbf{n}$ is equal to 1. Eventually, let O_i^v be the *visual occlusion* of $\mathbf{o}_i$ and set $= O_i^a$ as the *acoustic occlusion* of any object i (both in percent). With that, we define two identical functions $v_d(d_i) = a_d(d_i)$ in order to emulate degradation of visual sensor performance with increasing d_i. We further set up a function $a_w(\alpha_i)$, which encodes the dependency between an object's audibility and the corresponding α_i

$$g_d(d_i) = 1 - \frac{1}{1 + e^{-\frac{d_i - 10}{2}}}, \qquad a_w(\alpha_i) = e^{-0.5\frac{\alpha_i}{90}^2}$$

Using $g_d(d_i)$ as a proxy for both $v_d(d_i)$ and $a_d(d_i)$, it is herein assumed that an object's visibility and audibility remain acceptable, even for larger distances d_i. Further, our definition of $a_w(\alpha_i)$ postulates a relatively moderate decay in audibility for growing $|\alpha_i|$. In addition, let

$$v_o(O_i^v) = 1 - \frac{O_i^v}{100}, \qquad a_o(O_i^a) = 1 - 0.8\frac{O_i^a}{100}$$

reflect the degradation of visual and acoustic sensor performance in the presence of occlusion. Note that $v_o(O_i^v)$ renders the vision sensor powerless under full visual occlusion. Contrary, $a_o(O_i^a)$ remains ≥ 0.2 for $O_i^a \in [0..1]$. In the presence of obstacles, this 'residual audibility' approach allows for sound muffling, and simultaneously ensures reliable acoustic object detection. Based on the above, let

$$v_P(d_i, O_i^v) = v_d(d_i) \cdot v_o(O_i^v) \tag{1}$$

be the *visual perceptibility* of object i, whereas

$$a_P(d_i, O_i^a, \alpha_i) = a_d(d_i) \cdot a_o(O_i^a) \cdot a_w(\alpha_i) \tag{2}$$

is the *acoustic perceptibility* of $\hat{\mathbf{o}}_i$. With the above, we set up a degraded *visual labeling function* for each $\mathbf{o}_i$

$$v_L^j(l_i, v_P^i) = \begin{cases} 0.5 + \frac{0.5}{1+20.0\cdot e^{-(v_P^i - 0.5)}} & \text{if } l_i = k_j \\ 0.5 - \frac{0.5}{1+20.0\cdot e^{-(v_P^i - 0.5)}} & \text{if } l_i \neq k_j \end{cases} \tag{3}$$

where $k_j \in \mathcal{K}$ and v_P^i is a placeholder for $v_P(d_i, O_i^v)$. This function introduces significant uncertainty with respect to the true category of an object for low visual perceptibility values, that is, driving $v_P(d_i, O_i^v)$ towards zero, $v_L^j(l_i, v_P^i) \to 0.5$ for each category k_j. This behavior reflects the insight that in the physical world, a far-away, occluded object $\mathbf{o}_i$ would be hard to classify and could thus belong to any available category. Increasing the visual perceptibility, eq. 3 bifurcates, allowing object $\mathbf{o}_i$'s true category l_i to clearly dominate all other potential category estimates. That accommodates the assumption that inference in modern physical classifiers becomes highly reliable for well-behaving data, that is, objects that are close by and fully visible. Note that we set up an additional *auditory labeling function* by merely replacing $v_P(d_i, O_i^v)$ with $a_P(d_i, O_i^a, \alpha_i)$ in eq. 3.

3 Dynamic weighting module

To be able to take the reflex movements into consideration, a weighting approach has been developed. The dynamic weighting module (DW) represents a

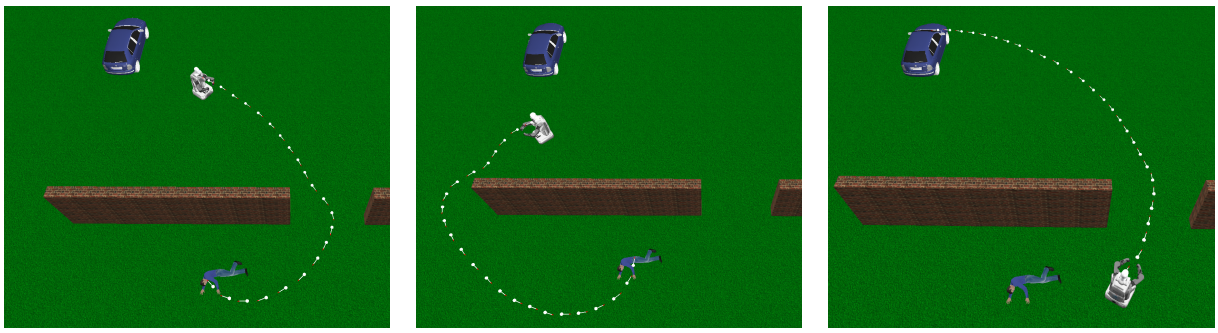

Figure 3: Path planning in the BEFT: three different motion patterns are planned, according to the robot's initial position. The planned paths are indicated by waypoints (white dots) enriched with additional orientation information (red-white arrows) that represents the platform's intended orientation at each waypoint.

simulated feedback loop that would biologically extend between the cochlea and the primary auditory cortex. The module mimics two biological phenomena that occur in auditory processing, namely: (i) Turn-to-reflex movements, and (ii) attentional filtering of sound sources. DW is based on the *congruency* of incoming auditory events. As in geometry or algebra, the congruency of two or more objects is a measure of the similarities of some features of these objects. In perception, congruency relates to perceptual and semantic features that link two objects (for example, seeing a dog and hearing it bark). If the reliability of an auditory stimulus is low, ambiguities due to lack of information or unexpected behavior of an object become likely. The Turn-to-reflex then possibly changes the robot's current task depending on the potential threat of an incoming stimulus. For the present work, DW is based on the labels of all relevant environmental objects. The dynamic weighting module is able to evaluate incoming stimuli according to their importance and eventually triggers a quick turn-to reaction or changes the robot's current task. The machine keeps an internal world representation of all *congruent environmental objects*, namely $\mathcal{C} = \{\hat{\mathbf{o}}_i^c\}, i = 0...N_C - 1$. For any newly detected object $\hat{\mathbf{o}}_i$, the corresponding weight will be set to 0.5. If $\hat{\mathbf{o}}_i \in \mathcal{C}$ and emits a novel sound stimulus, the DW module assigns less importance to the onset of this stimulus, potentially suppressing the turn-to-reflex, while quickly decreasing the corresponding object's weight. Weights are limited to lie in between 0 and 1, where 0 represents a highly congruent sound object and 1 indicates a highly incongruent and/or dangerous sound object. To that end, bifurcating logarithmic functions are used to model the dynamics of the machine's intended reaction:

$$w(t) = \begin{cases} \frac{log(t)}{2*log(n)} + c & \text{if } 1 \leq t < T_{max} \\ \frac{log(T_{max})}{2*log(n)} + c & \text{if } t \geq T_{max} \\ 2c - w_1(t) & \text{if } 1 \leq t < T_{max} \\ 2c - w_2(t) & \text{if } t \geq T_{max} \end{cases} \quad (4)$$

In eq. 4, let $c = (0.25, 0.5, 0.75)$ be the state of congruency, depending on the activity and the label of a given object. The higher c, the less congruent the new auditory object is. DW updates the weights in a discrete manner. The time interval between two time steps can be changed. T_{max} is a time threshold employed to counteract deadlock situations. For instance, if an object has the same weight for more than T_{max} time steps, the system reconsiders the object's congruency. Here, we choose $T_{max} = 10$. If any $\hat{\mathbf{o}}_i^c$ is suddenly detected as incongruent (for example, a person that was walking and is now yelling), the object's weight will abruptly increase to 0.75 and then further increase to 1. If the person stops yelling, that does not necessarily mean that the person is not in danger anymore. Thus, the corresponding object's weight slowly decreases to a residual value of 0.5.

3.1 Path planning

To realize and test active exploration strategies in the BEFT, autonomous path planning becomes a mandatory component. To that end, we decided to step away from full-fledged SLAM [34] for the moment, and instead set up an approach that combines *multi-hypotheses planning techniques* (loosely inspired by [35]) with well-known *potential field* methods. Be $\mathcal{F}$ a potential field assembled from the repulsive influences of all obstacles and the target's attraction well, where target coordinates will be provided by the dynamic weighting stage (see above). For the construction of the obstacles' repulsive potentials, the robot's simulated camera is used to construct an *obstacle map* $\mathcal{M}_O$ that stores the footprint coordinates of all obstacles that were so far faced by the moving platform. Blurring this map with a standard Gaussian kernel ($\sigma = 3.0$) followed by normalization yields the *potential map* $\mathcal{M}_P$, with $\mathcal{M}_P(\mathbf{x}) \in [0..1]$. With that, we define the *force map* $\mathcal{M}_F = -\nabla\mathcal{M}_P$. For the *target force map* $\mathcal{M}_T$, let $\mathcal{M}_D$ be the normalized *distance transformation map* generated for the target's position. From this follows $\mathcal{M}_T = -\nabla\mathcal{M}_D$. Given $\mathcal{M}_T$ and

$\mathcal{M}_O$, the robot starts path planning by emitting a set $\mathcal{H} = \{h_i\}, i = 0...N_P - 1$ of particles, with $N_P = 100$. Each particle corresponds to a certain *path hypothesis*. A standard gradient descent mechanism then drives the particles towards the target along the virtual lines of force. To escape local minima, and to avoid potential deadlock situations, each h_i is attributed with a random motion component that decays over time. This idea can loosely be compared to well-known *simulated annealing* strategies, and generates perceptually plausible paths in all our experiments. Due to the random component, particles likely traverse obstacles on different ways. We decide for the winning particle h_W based on path length and straightness. Given h_W, the virtual robot traces the corresponding path and approaches the given target while avoiding all already observed obstacles.

4 Results

Fig. 1 provides a convenient scaffold to state our results in the installation of an adequate 'feedback and communication architecture for [...] complex binaural and visual evaluation' (s. above): the proposed BEFT system integrates a custom-made virtual environment that allows for visualization and manual inspection of a wide range of scenarios that are likely to become relevant within the framework of TWO!EARS. The BEFT core is enabled to communicate with external system components via TCP/IP protocols in order to ensure maximum systemic flexibility. Together with the above degradation functions, the testbed's hierarchical system of experts allows to generate, analyze, and combine multimodal information, that is, visual and emulated acoustic features, to the end of forming high-level, symbolic representations of observed objects. Our approach incorporates baseline path planning techniques (see fig. 3) and schedules low- to high-level experts that prepare the ground for virtual bottom-up/top-down feedback mechanisms like active exploration. With the above, the main benefit gained from the proposed testbed architecture comes clear. In fact, our system will allow to integrate and test routines for complex feedback paths early in the TWO!EARS project, and enable our project partners to test their own ideas in a custom-tailored virtual environment. Due to this strategy, potential issues might be detected and eliminated long before final system assembly. Also, the BEFT will enable identification of environmental variates and labels that are definitely needed for reliable feedback. The proposed DW module showed promising results in first experiments performed in *MATLAB* ®. Fig. 4 yields results from a complex scenario that includes changing activities, for instance, a 'walking' person starting 'yelling'. The simulation also includes an initially harmless fire object that then endangers the 'walking' person. The title of the figure indicates the current task of the robot and the robot's current viewing direction, as computed by DW. The dashed lines symbolize the three different possible states of congruency (0.25, 0.5, and 0.75).

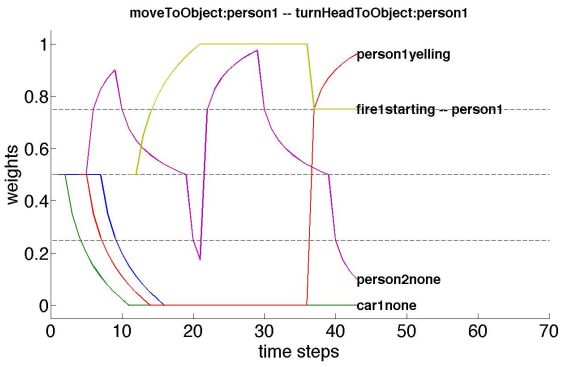

Figure 4: Dynamic weighting in *MATLAB* ®; refer to running text for details.

5 Conclusions

With the BEFT, we have created a versatile and extendible solution that allows for early feedback testing in virtual scenarios. While first feedback mechanisms (for instance, the dynamic weighting module) are currently integrated into our testbed, several things remain to be done in upcoming system generations: one systemic 'core' component, namely, the auralization client, is currently mimicked by plainly deriving acoustic estimates from degraded ground-truth scenario information. To eliminate this oversimplified approach, we intend to connect our testbed to a far more evolved auralization mechanism, such as the *Sound Scape Renderer* announced above. Further, the Petri nets used for expert scheduling need rework to accommodate more complex task plans. With respect to connectivity, our system's capabilities will have to be extended as well, in particular, more evolved TCP/IP channels are planned to allow for generic communication with external modules. Further, we intend to set up a specific *feedback class* that enables seamless *MATLAB* ® integration of our virtual testbed.

Acknowledgement

This research has been [partially] supported by EU FET grant TWO!EARS, ICT-618075, `www.twoears.eu`

References

[1] D. Calisi, A. Farinelli, L. Iocchi, D. Nardi, and R. La, "Multi-objective exploration and search for autonomous rescue robots," *Journal of Field Robotics*, vol. 24, no. 8-9, pp. 763–777, 2007.

[2] R. P. Würtz, ed., *Organic Computing*. Springer, 2008.

[3] C. Luo, A. P. Espinosa, D. Pranantha, and A. De Gloria, "Multi-robot search and rescue team," in *IEEE International Symposium on Safety, Security, and Rescue Robotics*, pp. 296–301, 2011.

[4] R. Stopforth, S. Holtzhausen, G. Bright, N. Tlale, and C. Kumile, "Robots for Search and Rescue Purposes in Urban and Underwater Environments - a survey and comparison," in *15th International Conference on Mechatronics and Machine Vision in Practice*, pp. 476 – 480, IEEE, 2008.

[5] A. Wolf, H. H. Choset, B. H. Brown, and R. W. Casciola, "Design and control of a mobile hyper-redundant urban search and rescue robot," *Advanced Robotics*, vol. 19, pp. 221–248, Jan. 2005.

[6] R. Teymourzadeh, R. N. Mahal, N. K. Shen, and K. W. Chan, "Adaptive Intelligent Spider Robot," in *IEEE Conference on Systems, Process & Control*, pp. 310–315, 2013.

[7] G. Campion and W. Chung, "Wheeled Robots," in *Springer Handbook of Robotics*, pp. 391–410, Springer, 2008.

[8] R. Sheh, "The Redback: A Low-Cost Advanced Mobility Robot for Education and Research," in *IEEE International Workshop on Safety, Security and Rescue Robotics*, 2006.

[9] W. Burgard, M. Moors, C. Stachniss, and F. E. Schneider, "Coordinated multi-robot exploration," *IEEE Transactions on Robotics*, vol. 21, pp. 376–386, June 2005.

[10] P. Michel, J. Chestnutt, S. Kagami, K. Nishiwaki, J. Kuffner, and T. Kanade, "GPU-accelerated real-time 3D tracking for humanoid locomotion and stair climbing," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 463–469, IEEE, Oct. 2007.

[11] "Springer Reference", "Definition of active exploration," 2014. `http://www.springerreference.com/docs/html/chapterdbid/179624.html`.

[12] S. B. Thrun and K. Möller, "Active Exploration in Dynamic Environments," in *Advances in Neural Information Processing Systems*, (San Mateo), pp. 531–538, Morgan Kaufmann, 1992.

[13] S. Ivaldi, S. M. Nguyen, N. Lyubova, A. Droniou, V. Padois, D. Filliat, and O. Sigaud, "Object learning through active exploration," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 1, pp. 56–72, 2013.

[14] A. Baranes and P.-Y. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robotics and Autonomous Systems*, vol. 61, pp. 49–73, Jan. 2013.

[15] Y. Sun, F. Gomez, and J. Schmidhuber, "Planning to Be Surprised : Optimal Bayesian Exploration in Dynamic Environments," in *4th International Conference on Artificial General Intelligence*, pp. 41–51, Springer, 2011.

[16] B. Goertzel, H. D. Garis, C. Pennachin, and N. Geisweiller, "OpenCogBot: Achieving Generally Intelligent Virtual Agent Control and Humanoid Robotics via Cognitive Synergy," in *Proceedings of the International Conference on Artificial Intelligence*, (Beijing), 2010.

[17] T. Xu, K. Kühnlenz, and M. Buss, "Autonomous Behavior-Based Switched Top-Down and Bottom-Up Visual Attention for Mobile Robots," *IEEE Transactions on Robotics*, vol. 26, no. 5, pp. 947–954, 2010.

[18] L. J. Manso, P. Bustos, and P. Bachiller, "Multi-cue Visual Obstacle Detection for Mobile Robots," *Journal of Physical Agents*, vol. 4, no. 1, pp. 3–10, 2010.

[19] A. Sanfeliu, "Robot Vision for autonomous object learning and tracking," in *8th Iberoamerican Congress on Pattern Recognition*, vol. 2905, (Havana), pp. 17–28, Springer Berlin Heidelberg, 2003.

[20] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments," *The International Journal of Robotics Research*, vol. 31, pp. 647–663, Feb. 2012.

[21] H. Zhang, C. Reardon, and L. E. Parker, "Real-time multiple human perception with color-depth cameras on a mobile robot.," *IEEE Transactions on cybernetics*, vol. 43, pp. 1429–1441, Oct. 2013.

[22] Y. V. S. Chan, *Audio-Visual Sensor Fusion For Robotic Source Localisation*. PhD thesis, University of Sydney, 2008.

[23] S.-W. Kim, J.-Y. Lee, D. Kim, B.-J. You, and N. L. Doh, "Human Localization based on the fusion of Vision and Sound System," in *8th International Conference on Ubiquitous Robots and Ambient Intelligence*, pp. 495–498, IEEE, 2011.

[24] A. Raman, J. Buck, D. Rogers, J. Drabner, M. Sari, and M. Goldberg, "OGRE (Object-Oriented Graphics Rendering Engine)," 2014. `http://www.ogre3d.org/`.

[25] M. Geier and S. Spors, "Spatial Audio Reproduction with the SoundScape Renderer," in *27th Tonmeistertagung, VDT International Convention*, 2012.

[26] Willow Garage, "The PR2 robot," 2014. `http://www.willowgarage.com/pages/pr2/overview`.

[27] G.R.A.S. Sound&Vibration, "KEMAR-rejuvenated," 2014. `http://kemar.us/`.

[28] G. Enzner, C. Antweiler, and S. Spors, "Trends in Acquisition of Individual Head-Related Transfer Functions," in *Modern Acoustics and Signal Processing - The Technology of Binaural Listening* (J. Blauert, ed.), pp. 57–92, Springer Berlin Heidelberg, 2013.

[29] J. Hsu, E. Berger, and A. Hendrix, "PR2 description," 2014. `http://wiki.ros.org/pr2_description`.

[30] M. Bastioni, "MakeHuman^{TM} - Open Source tool for making 3D characters," 2014.

[31] BlenderFoundation, "Blender modeling software," 2014. `http://www.blender.org/`.

[32] J. Blauert, D. Kolossa, K. Obermayer, and K. Adiloğlu, "Further Challenges and the Road Ahead," in *Modern Acoustics and Signal Processing - The Technology of Binaural Listening* (J. Blauert, ed.), pp. 477–501, Springer Berlin Heidelberg, 2013.

[33] T. Murata, "Petri Nets : Properties , Analysis and Applikations," *Proceedings of the IEE*, vol. 77, no. 4, pp. 541–580, 1989.

[34] W. Burgard, C. Stachniss, G. Grisetti, B. Steder, R. Kummerle, C. Dornhege, M. Ruhnke, A. Kleiner, and J. D. Tardos, "A comparison of SLAM algorithms based on a graph of relations," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2089–2095, Oct. 2009.

[35] M. Oliveira, V. Santos, and A. Sappa, "Short term path planning using a multiple hypothesis evaluation approach for an autonomous driving competition," in *4th Workshop on Planning,Perception and Navigation for Intelligent Vehicles*, IEEE, 2012.