11

A Multi-Modal Panoramic Attentional Model for Robots and Applications

Ravi Sarvadevabhatla and Victor Ng-Thow-Hing
Honda Research Institute USA, 425 National Ave Suite 100,Mountain View CA 94043 USA

1. Introduction

Humanoid robots are becoming increasingly competent in perception of their surroundings and in providing intelligent responses to worldly events. A popular paradigm to realize such responses is the idea of attention itself. There are two important aspects of attention in the context of humanoid robots. First, *perception* describes how to design the sensory system to filter out useful salient features in the sensory field and perform subsequent higher level processing to perform tasks such as face recognition. Second, the *behavioral response* defines how the humanoid should act when it encounters the salient features. A model of attention enables the humanoid to achieve a semblance of liveliness that goes beyond exhibiting a mechanized repertoire of responses. It also facilitates progress in realizing models of higher-level cognitive processes such as having people direct the robot's attention to a specific target stimulus(Cynthia et al., 2001).

Studies indicate that humans employ attention as a mechanism for preventing sensory overload(Tsotsos et al., 2005),(Komatsu, 1994) – a finding which is relevant to robotics given that information bandwidth is often a concern. The neurobiologically inspired models of Itti (Tsotsos et al., 2005), initially developed for modeling visual attention have been improved(Dhavale et al., 2003) and their scope has been broadened to include even auditory modes of attention(Kayser et al., 2008). Such models have formed the basis of multi-modal attention mechanisms in (humanoid) robots(Maragos, 2008),(Rapantzikos, 2007).

Typical implementations of visual attention mechanisms employ a bottom-up processing of camera images to arrive at the so-called "saliency map", which encodes the unconstrained salience of the scene. Salient regions identified from saliency map are processed by higher-level modules such as object and face recognition. The results of these modules are then used as referential entities for the task at hand(e.g. acknowledging a familiar face, noting the location of a recognized object). Building upon the recent additions to Itti's original model(Tsotsos et al., 2005), some implementations also use top-down control mechanisms to constrain the salience(Cynthia et al., 2001),(Navalpakkam and Itti, 2005),(Moren et al., 2008).

In most of the implementations, the cameras are held fixed, simplifying processing and consequent attention mechanism modeling. However, this restricts the visual scope of attention, particularly in situations when the robot has to interact with multiple people who may be spread beyond its limited field-of-view. Moreover, they may choose to advertise their presence through a non-visual modality such as speech utterances.

Attempts to overcome this situation lead naturally to the idea of widening the visual scope and therefore, to the idea of a *panoramic* attention. In most of the implementations which

utilize a panorama-like model (Kayama et al., 1998),(Nickel and Stiefelhagen, 2007), the panoramic region is discretized into addressable regions. While this ensures a complete coverage of the humanoid's field of view, it imposes high storage requirements. Most regions of the panorama remain unattended, particularly when the scene is static in nature. Even when dynamic elements are present(e.g. moving people), the corresponding regions require attention for a limited amount of time before higher-level tasks compel the system to direct its attention to other parts of panorama(Nickel and Stiefelhagen, 2007).

These limitations can be addressed by employing a *multi-modal panoramic attention model* – the topic of this chapter. In its basic mode, it operates on an egocentric panorama which spans the pan and tilt ranges of the humanoid's head cameras(Nickel and Stiefelhagen, 2007). As a baseline characteristic, the model naturally selects regions which can be deemed interesting for cognitively higher-level modules performing face detection, object recognition, sudden motion estimation etc. However, saliencies are maintained only for cognitively prominent entities (e.g. faces, objects, interesting or unusual motion phenomena). This frees us from considerations of storage structures and associated processing that are present in image pixel-based panoramic extensions of the traditional attention model(Kayama et al., 1998),(Ruesch et al., 2008). Also, the emphasis is not merely on obtaining a sparse representation in terms of storage as has been done in previous work. One of the objectives is also to assign and manipulate the semantics of sparsely represented entities in an entity-specific fashion. This chapter describes how this can be achieved with the panoramic attention model.

The panoramic attention model has an idle mode driven by an idle-motion policy which creates a superficial impression that the humanoid is idly looking about its surroundings. Internally, in the course of these idle motions, the humanoid's cameras span the entire panorama and register *incidental observations* such as identities of people it comes across or objects present in gaze directions it looks at. Thus, the idle-motion behavior imparts the humanoid with a human-like liveliness while it concurrently notes details of surroundings. The associated information may be later accessed when needed for a future task involving references to such entities i.e. the humanoid can immediately attend to the task bypassing the preparatory search for them. The active mode of the panoramic attention model is triggered by top-level tasks and triggers. In this mode, it responds in a task-specific manner (e.g. tracking a known person). Another significant contribution from the model is the notion of *cognitive panoramic habituation*. Entities registered in the panorama do not enjoy a permanent existence. Instead, their lifetimes are regulated by entity-specific persistence models(e.g. isolated objects tend to be more persistent than people who are likely to move about). This habituation mechanism enables the memories of entities in the panorama to fade away, thereby creating a human-like attentional effect. The memories associated with a panoramically registered entity are refreshed when the entity is referenced by top-down commands.

With the panoramic attention model, out-of-scene speakers can also be handled. The humanoid robot employed(Honda, 2000) uses a 2-microphone array which records audio from the environment. The audio signals are processed to perform localization, thus determining which direction speech is coming from, as well as source-specific attributes such as pitch and amplitude. In particular, the localization information is mapped onto the panoramic framework. Subsequent sections shall describe how audio information is utilized.

At this point, it is pertinent to point out that the panoramic attention model was designed for applicability across a variety of interaction situations and multi-humanoid

platforms. Therefore, having the model components,particularly the interfaces with sensors and actuators, operate in a modular fashion becomes important. Apart from technical considerations, one of the design goals was also to provide a certain level of transparency and user-friendly interface to non-technical users, who may not wish or need to understand the working details of the model. As described later, the model is augmented by an intuitive panoramic graphical user interface(GUI) which mirrors the model helps avoid cognitive dissonance that arises when dealing with the traditional, flat 2-d displays – a feature that was particularly valuable for operators using the interface in Wizard-Of-Oz like fashion.
The chapter will first review previous and related work in attention, particularly the panoramic incarnations. The details of the panoramic attention model are provided next, followed by a description of experimental results highlighting the effectiveness of the cognitive filtering aspect in the model. Interactive application scenarios (multi-tasking attentive behavior, Wizard-of-Oz interface, personalized human-robot interaction) describing the utility of panoramic attention model are presented next. The chapter concludes by discussing the implications of this model and planned extensions for future.

2. Related work

The basic premise of an attention model is the ability to filter out salient information from the raw sensor stream (camera frames, auditory input recording) of the robot(Tsotsos et al., 2005). In most implementations, the model introduced by Itti et al. (Itti et al., 1998) forms the basic framework for the bottom-up attention mechanism where low-level features such as intensity, color, and edge orientations are combined to create a *saliency map*. The areas of high saliency are targeted as candidates for gaze shifts (Cynthia et al., 2001; Dhavale et al., 2003; Ruesch et al., 2008; Koene et al., 2007). In some cases, these gaze points are attended in a biologically-inspired fashion using foveated vision systems(Sobh et al., 2008). In order to decide how to combine the various weights of features to create the saliency map, a top-down attention modulation process is prescribed that assigns weights generally based on task-preference(Cynthia et al., 2001). This top-down weighting can be flexibly adjusted as needed(Moren et al., 2008) according to the robot's current task mode.
The term panoramic attention has also been used to describe cognitive awareness that is both non-selective and peripheral, without a specific focus(Shear and Varela, 1999). There is evidence that the brain partially maintains an internal model of panoramic attention, the impairment of which results in individuals showing neglect of attention-arousing activities occurring in their surrounding environment(Halligan and Wade, 2005).
Although mentioned as a useful capability(Cynthia et al., 2001; Kayama et al., 1998), there have been relatively few forays into panorama-like representations that represent the scene beyond the immediate field of view. The term panoramic attention is usually confined to using visual panorama as an extension of the traditional fixed field-of-view representations (Kayama et al., 1998) or extension of the saliency map(Bur et al., 2006; Stiehl et al., 2008). The work of (Ruesch et al., 2008) uses an egocentric saliency map to represent, fuse and display multi-modal saliencies. Similar to the work described in this chapter, they ensure that salient regions decay. However, this mechanism operates directly on raw saliency values of discretized regions mapped onto the ego-sphere. In contrast, the cognitive decay mechanism (described in a subsequent section) works on sparsely represented higher-level entities such as faces. (Nickel and Stiefelhagen, 2007) employ a multi-modal attention map that spans the discretized space of pan/tilt positions of the camera head in order to store particles derived from visual and acoustic sensor events. Similar to the model presented in the chapter,

they exploit the information in their panoramic attention map to decide on gaze behavior for tracking existing people or for looking at novel stimuli to discover new people. The idea of a 'stale panorama' from (Stiehl et al., 2008) served as a motivating aspect for the present work.However, in their implementation only the raw video frames are used to create a panoramic image for a teleoperation implementation. The model presented in the chapter requires less storage because only the high-level semantic information from the visual field is extracted and stored such as identification and location of entities in the visual field. This is also important in that these semantic features can be used by the robot for decision-making, whereas further processing would need to be done with a raw image panorama. As part of their humanoid active vision system, (Koene et al., 2007) refer to a short term memory module that stores the location of perceived objects when they are outside the current field of view. They use this mechanism to associate sound with previous visual events, especially since audio may come from a source not in view.

In many of the panorama-like models mentioned above, the user interface does not mirror the world as perceived by the humanoid robot. The benefits of matching interface displays and controls to human mental models include reductions in mental transformations of information, faster learning and reduced cognitive load (Macedo et al., 1999) – a factor which inspired the design of the application interfaces described along with the model. As described in Section 5.1, the user interface mirrors the panoramic portion of the attention model, thereby minimizing cognitive dissonance. In the interest of focus and space, the numerous references to various Wizard-of-Oz systems and robot control interfaces will not be discussed.

In general, the aim for panoramic attention model is to fulfill a broader role than some of the aforementioned approaches. In particular, the sensor filtering characteristics of the bottom-up attention modules are combined with higher level spatial and semantic information that can then be exploited to by a behavior module for the robot. The semantic knowledge of an object allows refined modules to model the spatio-temporal behavior of objects perceived.

3. The panoramic attention model

A simple three-layer attention model forms the backbone of the panoramic attention model with each successive layer processing the outputs of its predecessor. Higher layers process data with increasing cognitive demand than the lower ones (Figure 1). To begin with, the visual portion of the attention model is described, followed by the audio portion.

3.1 Visual attention

Low-level visual attention: The traditional role of low-level visual attention is to rapidly determine the most salience-worthy saccade[1] locations for a given scene in a context-independent fashion.

The lowest among the layers, this processes input sensor data, i.e. color video frames streaming in from the humanoid's cameras. The processing for low-level layer happens as outlined in (Itti et al., 1998). Each input color frame is processed to arrive at the target ROI(Region of Interest)s. To improve the color contrast of incoming frames, Contrast-Limited Adaptive Histogram Equalization(Pizer et al., 1990) is applied on them. The contrast-adjusted color frame and its gray-scale version are used to generate various feature maps. The generic

[1] According to `http://en.wikipedia.org/wiki/Saccade`, a saccade is a fast movement of an eye, *head* or other part of an animal's body or device. The term saccade in this chapter refers to head movement since the humanoid used does not have a foveated vision control mechanism.

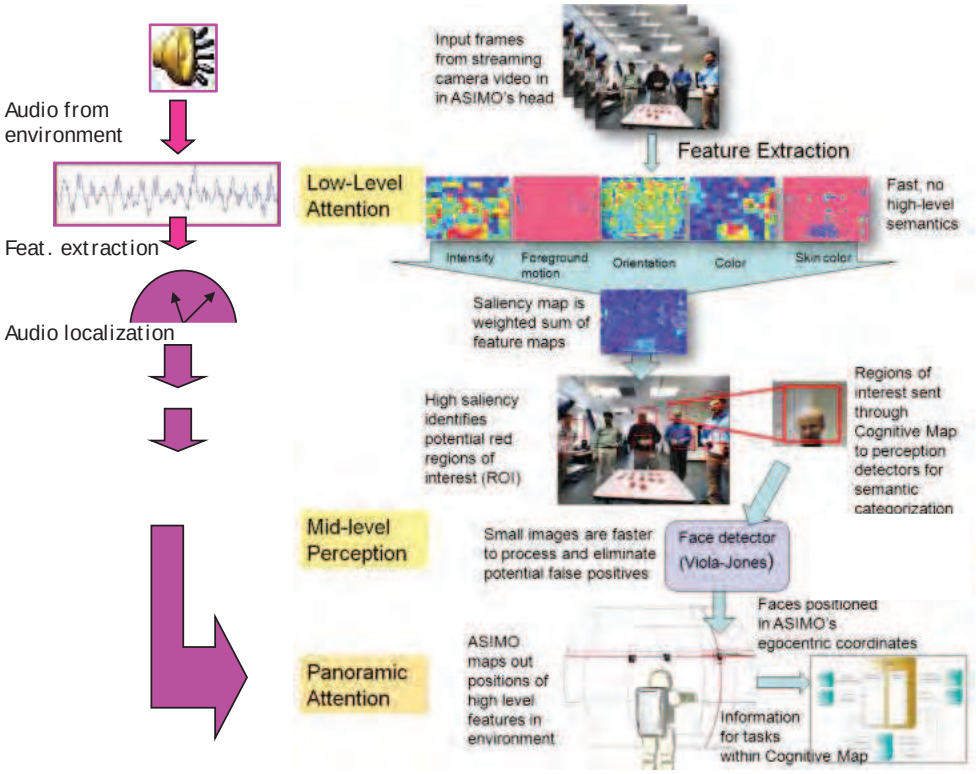


Fig. 1. Three layer panoramic attention model

procedure is to generate a Gaussian pyramid from input frame, perform center-surround differencing on select pyramid level pairs and add up the differences and finally normalize them – for details on feature map generation, refer to (Itti et al., 1998)[2]. In the current implementation, Color, Intensity, Orientation, Skin, and Motion features are used.

The base-level intensity map is derived from the corresponding channel of the HSI color space. The Color and Orientation feature maps are again constructed as suggested in (Itti et al., 1998). For Motion map, an adaptive foreground estimation method(McFarlane and Schofield, 1995) is used, which works well even in the presence of variable rate background motion. For Skin map, samples of skin segments were collected from training images and estimate the probability of skin pixels(Kakumanua et al., 2006). Another option for Skin map is to threshold the ratio of red and green channels of the RGB image and perform connected-component analysis on the result to obtain the feature-map level ROIs. The current implementation bypasses the pyramid reconstruction and interpolates the Motion and Skin feature maps to the level of other feature maps before forming the final saliency map.

[2] (Itti et al., 1998) refer to feature maps as conspicuity maps.

The saliency map is computed as a weighted addition of Feature Maps, with the weight proportional to the importance of the corresponding map. The following weights: $w_{color} = 0.1, w_{intensity} = 0.1, w_{orientation} = 0.1, w_{skin} = 0.2, w_{motion} = 0.5$ are used. The weights suggest relative importance to motion and skin color cues making the attention system more sensitive to regions that may contain faces as needed in human interaction scenarios. To generate the attentional saccade locations, a habituation/inhibition mechanism is applied on the saliency map(Ruesch et al., 2008). The Region of Interest (ROI)s are obtained as 2-D masks centered around each attention saccade location. Feature maps are generated from features of each frame and combined to produce a saliency map. A temporally dynamic habituation-inhibition mechanism is then applied to the saliency map to produce a sequence of saccade locations. The layer finally outputs Regions of Interest (ROI), which correspond to 2-D masks centered on eye-gaze locations.

Human-robot interaction scenarios may also require proximity-related reasoning(e.g. relative positions of people surrounding the robot). For this reason, the depth map associated with the RGB image is also processed. However, instead of processing the entire depth map, the depth map regions corresponding to the ROIs obtained are processed and a depth measurement relative to the robot's location is estimated for each ROI. Since the proximity concerning people is of interest, the Skin feature map is utilized. Connected component analysis is performed on the Skin feature map. For each component, the depth values corresponding to the location of each component's pixels are thresholded (with a lower and upper threshold to eliminate outliers) to obtain valid depth values. A histogram of such valid depth values is estimated for each component. The location of the histogram bin with the largest bin count is assigned as the depth value of the ROI.

Mid-level perception: This layer serves as a filter for the outputs of low-level attention. This layer forms the proper location for detectors and recognizers such as those employed for face/object detection and localization. Outputs from the low-level attention – the ROIs mentioned above – are independent of context. Effectively, the mid-level perception grounds and categorizes the ROIs into higher-level entities such as faces and objects, which are then passed on to the panoramic attention layer.

The mid-level perception contains face/object localization and recognition modules. Indeed, it is only at mid-level perception that some semblance of identity arises(Tsotsos et al., 2005). As mentioned before, one role of mid-level perception layer is to serve as a filter for incoming ROIs. ROIs provide candidate regions from the original image for the face module. In Section 4, the performance of mid-level perception layer is examined in its filtering role with an analysis of the reasons behind the achieved performance.

Panoramic attention layer: This layer manages the spatial and temporal evolution of entities detected by the robot's visual attention over its entire field of view, including the additional range obtained from the pan and tilt angle ranges in the camera head. To accomplish this, it registers entities from the mid-level perception onto a *cognitive*[3] panorama. It is also at this layer that auditory information is utilized to perform the aforementioned entity registration in a multi-modal fashion. The persistence of registered entities is managed in the course of updates to the panorama. External modules can communicate with the panoramic attention module and the information for their specialized purposes. In addition, an idle-motion policy

[3] The word *cognitive* here denotes the semantic information stored in this panorama and also serves to distinguish from the usual definition of panorama as a large field of view composite image stitched from smaller images.

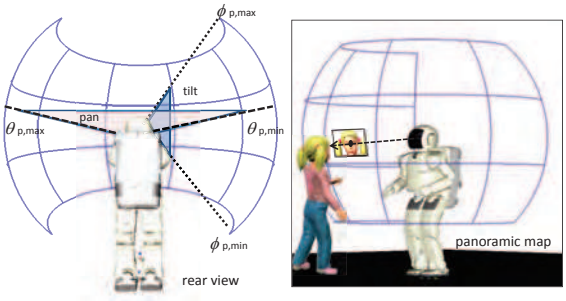


Fig. 2. Left: Panoramic map centered on ego-centric view from camera. Right: Given knowledge of the robot's head position and orientation and the 2-D image, it can be deduced that the person is somewhere along the line of sight connecting the camera origin through the entity's position in the 2-D image.

determines the current portion of the panorama that should be attended to, in the absence of any top-level controls.

The panoramic attention layer places the information about each entity obtained from the mid-level perception layer into the robot's ego-centric spherical coordinate system centered about its head (Figure 2). This information can then be accessible by behavior modules performing tasks within the robot system. In contrast to the lower layers which receive low-level sensor or image information, the input to the panoramic attention layer consists of higher level semantic features, such as face or object positions.

The main purpose of the panoramic attention layer is to expand environmental awareness beyond the robot's immediate field of view. It encompasses the total range of vision given the robot's head motion and the field of view of its cameras. Although in the current implementation only 2-D information is stored, the panoramic map can be easily augmented with three dimensional information given depth information obtained through stereo images or time-of-flight sensors. Nevertheless, given 2-D positions and the current joint angles of the head, some spatially-aware tasks are achievable. For example, in Figure 2, the position of the girl's face lies somewhere along the line of sight cast from the camera origin to the image's 2-D projection of face on the panoramic surface. If depth information is available, the distance along this ray can be esolved. Although items in the panoramic map cannot be localized for manipulation or navigation tasks without depth information, there is enough information to conduct many useful interaction tasks such as directed gaze because only relative directional information from the robot is required. If entities are beyond arm's reach of the robot, pointing operations can also be performed.

To accomplish these useful interaction tasks, an egocentric-based coordinate system (Figure 2, Left) is used, where the origin is localized at the robot's center of rotation of its head. Since all entities in the panoramic map are stored in the camera coordinate system and the configuration of the camera on the robot's body is known, the relative directional information of an entity with respect to its head can be reconstructed. If depth information is known, the relative position can also be estimated. Raw image data from the cameras are rectified to eliminate spherical distortion before being sent to the mid-level perception detectors. Consequently, there is no need to compensate for intrinsic camera parameters for the incoming image coordinates sent to the panoramic attention layer.

As the robot detects new entity locations, they are stored in the corresponding positions in the panorama. The positions are updated depending on the information provided by the mid-level perception modules. For example, if the mid-level modules perform tracking, the new positions simply replace the existing entries in the panoramic map. If no tracking is done, tracking behavior is built into the panoramic map using nearest-neighbour updates by treating the observation events as a time series of samples.
As the observations made by the panoramic map are made in an opportunistic manner (no explicit instructions to seek out or observe the entity), this allows certain commands to be performed with minimal preparation time. For example, when an instruction to look or point is given for an object already stored in the map, the action is carried out immediately without first needing to find the object, even in situations where the entity is off-screen.

3.2 Auditory attention

The audio signal from the environment is captured via the microphone array situated on the robot. Using the difference in time taken for the audio signal to travel to the microphones in the array, the azimuthal location of sounds(i.e. direction of sound as referenced in a horizontal plane) is localized(see Figure 1). This process of localization can be considered as selection of salient sounds(speech, specifically), analogous to the task of determining ROIs in low-level visual attention processing. In addition, pitch and amplitude features are extracted from the audio signal. The localization information, along with these extracted features, is combined at the panoramic layer of the model thereby making the process of entity registration a multi-modal affair. At the panoramic level, this incoming information is analyzed to identify the number of active speakers in the environment. The design choice of combining audio information with the visual in order to perform entity registration at the panoramic level can be considered as a late-binding model,wherein the modalities(audio and visual) are processed independently and in parallel to eventually combine at cognitively high levels where modality-specific information tends to be fairly grounded.
At the panoramic level, the audio-related information is received in the form of *sound sample messages*. Each message is associated with an angle corresponding to the azimuthal direction of the origin of the sound from the environment. In order to identify the number of speakers, the messages are clustered in real-time using the azimuthal angle as the clustering attribute, each cluster corresponding to a speaker and the center of the cluster corresponding to the azimuthal directional location of the speaker(Guedalia et al., 1998). During interaction, it is quite natural for the speakers to move about in the environment and this includes entirely disappearing from the zone of interaction. Therefore, the created clusters, which correspond to *active* speakers, do not persist indefinitely. Each cluster C_i is associated with an activation α_i modelled as an exponential decay function:

$$\alpha_i = \exp(-t/\tau_i), \tag{1}$$

where t represents the elapsed time since the last sample was added to the cluster and τ_i is a time-constant used to control the rate of decay.
As the cluster C_i is created, it is initially assigned an activation $\alpha_i = 1$. The activation value then decays according to the above equation. If a new sound sample falls into the cluster, the activation value is reset to 1. Otherwise, if the activation value falls below a threshold, the cluster is removed.

4. Experimental assessment of using ROIs

An immediate consequence of the panoramic attention model is that it doubles up as a filter for part-based, localized detection and recognition modules which reside at the mid-level perceptual layer. As mentioned previously, this mechanism decreases both the processing time and false positive rate for these modules. To verify this, these two quantities were measured in videos captured by the humanoid. The video data was collected with varying number of people and their activities (e.g. standing still, gesturing, conversing with each other, walking). The results in Table 1 validate the filtering role of the model[4]. The explanation for decrease in processing time is obvious. The consequence of the decreased processing time becomes significant in that additional time can be devoted for detailed processing of ROIs(e.g. determining facial expressions) without loss in throughput. It must also be noted that the decrease in false-positive rate is not simply a consequence of decrease in size of processing region. The low-level attention model described in the previous section provides good candidate image regions(ROIs) which have been selected using multiple cues such as color, intensity, skin and motion-regions. In particular, the skin and motion related feature maps offer valuable information about human presence. This, combined with the appropriate weighting of feature maps, enables the selection of ROIs which tend to contain good face candidates. On the other hand, the inability to exploit the aforementioned human-related cues contributes to larger false positive rates when face detection is performed on entire image by face detection algorithms which cannot utilize real-time human presence cues. Arguably, the model could fail to attend all of the entities present in the time-span of processing a single frame(thereby resulting in a non-trivial false negative rate), but this is a trade-off between the number of interest regions considered (consequently, processing time) and the rate at which the inhibition-habituation occurs over time.

Region size	Processing time	False positive rate
Full-frame (1024x768)	1204.71 milli-sec	4.74%
ROIs (200x200)	96.26 milli-sec	1.24%

Table 1. A quantitative measurement of filtering role played by the attention model

5. Applications

Once the panoramic attention framework is established, its information can be used to model various behaviors. In particular, models to direct the robot's gaze under various situations such as idle moments, encountering a new individual and following conversations between humans will be described in this chapter. Studies have shown that a robot's gaze cues can manipulate people's behavior and the roles they see themselves in during interaction with robots and other humans. For example, gaze can indicate strong turn-yielding cues to participants in a conversation (Mutlu et al., 2009). In idle moments when the robot is not conducting any primary tasks, the lack of visible activity can lead people to believe the robot is no longer functional. For this reason, small head motions are sometimes introduced in the robot's behavior to indicate liveliness. In (Nozawa et al., 2004), head motions were randomly perturbed to give it visual expression while it talked to compensate for the robot's lack of a mouth.

[4] The full-frame results were obtained using the OpenCV implementation of the Viola-Jones face detector(Viola and Jones, 2004)

Fig. 3. Panoramic attention in action: In the absence of attention-triggers, the robot looks about idly with the gaze locations generated by idle-policy model. The associated gaze direction probabilities can be seen at the bottom of the panorama in green as percentages. In the left figure, the robot has just seen a familiar face so it registers identity and location on panorama. When it observes a second person(right figure), it repeats the registration process.

Rather than play pre-recorded or purely random motions, the panoramic attention framework is used in combination with gaze control to let the robot actively map out its surrounding environment during its idle moments. This forms the basis of the panoramic layer's idle-motion policy. The panorama is first subdivided along its horizontal camera pan angle range $[\theta_{min}, \theta_{max}]$ into n bins of width w where:

$$w = \frac{\theta_{max} - \theta_{min}}{n} \tag{2}$$

and the bounds of bin i are defined as:

$$(\theta_{min} + i * w, \theta_{min} + (i+1) * w), i = 0, \cdots, n-1. \tag{3}$$

A frequency histogram can be built from the number of gaze occurrences that fall into each bin with f_i corresponding to the frequency of bin i. It is desirable to have the robot select new gaze directions based on areas it has visited infrequently to ensure it is adequately sampling its entire viewing range. Consequently, the probability of visiting bin i is defined as:

$$p_i = \frac{\max_j f_j - f_i}{\max_j f_j * n - f_{total}} \tag{4}$$

where f_{total} is the total number of gaze hits. From Equation 4, a bin that receives many gaze hits is less likely to be selected while those bins that have no current hits are very likely to be seen next. This strategy eventually allows the robot to map out entities in its surrounding environment while it is idle. (Refer to Figures 3 and 5 where the probabilities are marked at the bottom of the panorama in green). A lively-looking humanoid also results as a convenient by-product of this idle-policy.

In situations where entities appear in the robot's view (such as faces), their positions are noted within the panoramic map (Figure 4). If the face is recognized as a new person, the robot can respond to the person's presence(e.g. by greeting him/her) and subsequently tracks the changing position of the person via updates from the mid-level attention layer. This can be done for more than one person in a surreptitious manner while the robot is performing other tasks.

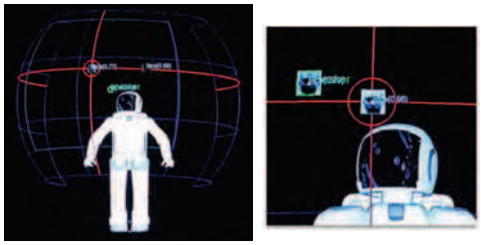


Fig. 4. Multiple faces appear in panoramic attention system (red cross-hairs represent current focus)

In situations where the robot hears the name of a familiar person during a conversation, if the person's location is already in the panoramic map, the robot can redirect its gaze towards that person, providing a strong visible reminder to the person that conversation was directed at him or her. For example, in Figure 3(left), the robot first recognizes a person and registers his location, then its attention is directed to a second person(Figure 3(right)). When the latter asks the former a question(Figure 5(left)), the robot immediately knows the location of the person being referenced and can look back at him without searching(Figure 5(right)). In case the

Fig. 5. The robot follows the conversation between them. When it hears a referential phrase (e.g. name being called)(left), it proceeds to look at the referent and participate in the conversation appropriately(right).

person being referenced has moved in the meantime, the robot commences the search from the last known stable position. If the person can be tracked within the field of view of the first person, the robot skips the search phase. Either way, this strategy is more efficient than a full-blown panorama-wide search.

When observations of faces appear in the system, they are assigned an activation signal initially set to $\alpha = 1$. As time progresses, the signal decays according to the exponential decay function:

$$\alpha = \exp(-t/\tau), \tag{5}$$

where t refers to time passed since the initial observation and the decay rate is set to τ, with smaller values of τ causing faster decay.

This activation signal is used to provide a finite bound on the time the observation stays in the panoramic map. If a new observation of the entity is made, the activation signal is reset to 1. The decay rate can be assigned different values based on entity type. For example, static

and isolated objects can be given a slow decay rate while moving people can be assigned faster decay rates. The activation signal can also indicate how stable an observation is. For example, a rapidly moving face will have its activation constantly being reset to one. When encountering people in the scenario mentioned previously, they were often observed in motion before coming to a stop. When the robot's head was initially set to look at the person on the first observation event, the robot would often look at the position where the person was just moments before. However, if the robot is directed to look at the person only after the activation signal has reduced to a lower value threshold, the robot will look at a person only after his or her position stays stable long enough for the activation signal to decay to the threshold set. This behavior is called *cognitive panoramic habituation* – a panoramic analogue of the habituation-inhibition model in traditional attention architectures.
When the robot moves its head, a *head-motion-pending* message is broadcast to temporarily disable the low-level attention system and prevent incorrect motion-detected events caused by the robot's own head motion. When head motions have stopped, a *head-motion-stopped* message is broadcast to reactivate the low-level attention system. This mechanism also informs the high level panoramic attention layer to ignore incoming face position events since locating its position within the 3-D panoramic surface would be difficult because the robot's view direction is in flux. This strategy is supported by studies related to saccade suppression in the presence of large scene changes(Bridgeman et al., 1975).

5.1 Other applications

When used simultaneously with multiple modalities, this combined information can create a better estimate of the robot's environment. Since the panoramic attention map encompasses the robot's entire sensory range of motion beyond the immediate visual field, entities not immediately in camera range can still be tracked. The ability for the robot to obtain a greater sensory awareness of its environment is utilized in the following applications:

5.1.1 Wizard-of-Oz interface

In addition to providing monitoring functions, the panoramic map can be used an interactive interface for allowing a robot operator to directly select targets in the robot's field of view for specifying pointing or gaze directions (Figure 6). The two-dimensional image coordinates of entities within the active camera view are mapped onto the three-dimensional panoramic view surface, producing three dimensional directional information. This information enables a robot or any other device with a pan/tilt mechanism to direct their gaze or point their arm at the entity in its environment using its egocentric frame of reference(Sarvadevabhatla et al., 2010). Two important features of the interface are evident from Figure 6 – the interface mirrors the world as seen by the robot and the semi-transparent, movable operator GUI controls ensure a satisfactory trade-off between viewing the world and operating the interface simultaneously.

5.1.2 Multi-tasking attentive behavior

The panoramic attention model can be extended to manage other modalities such as motion. This allows a person to get a robot's attention by waving their hands. This ability is important in multi-tasking situations where a robot may be focused on a task. In Figure 7, a person waves his hand to get the robot's attention while the robot is actively playing a card game with a different person. The attention-seeking person may either shout a greeting (possibly off-camera) or wave their hands in the robot's visual field. Once the robot directs its gaze to

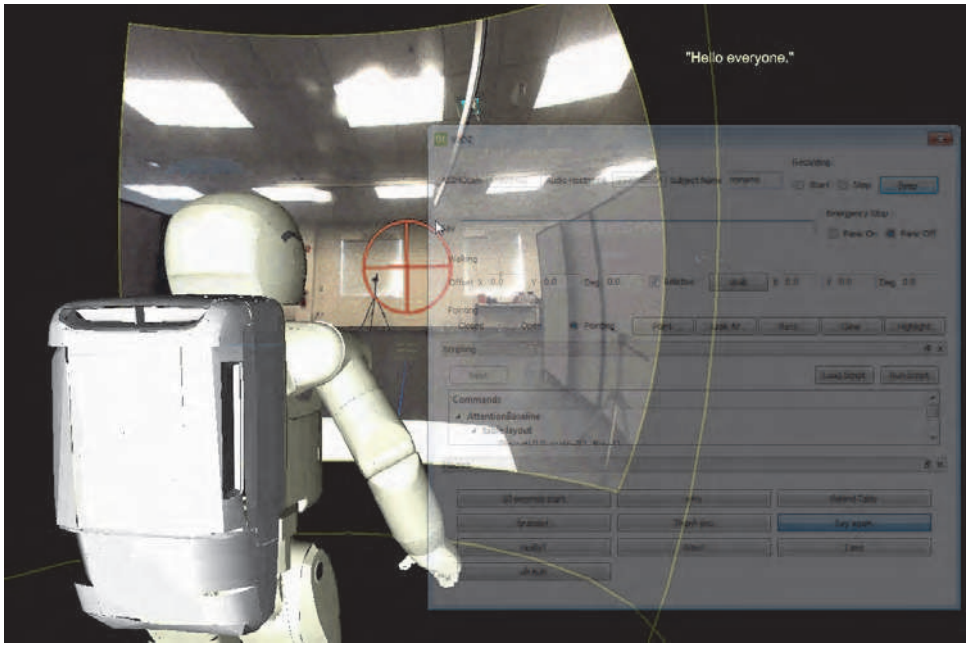


Fig. 6. Wizard-of-OZ application for direct selection of gaze and pointing targets

the source of the sensory disturbance, face detectors automatically locate the face to allow the robot to shift its gaze directly to the person's face.

5.1.3 Multi-speaker logging

The panoramic map can be used to log information and assign these as properties to entities in its environment. For example, we have actively kept track of the amount of speaker participation in group scenarios and attributed spoken utterances to the likely speaker based on the location of the sound utterance (Figure 8). Utilizing multi-modal information, it can be confirmed if a sound utterance coincides with a face appearing in the same location to avoid false positive identification of people by non-human sound sources in the environment. This mechanism has been used to allow the robot to monitor relative activity from a group of participants in a conversation.

6. Discussion and future work

The low-level attention layer utilized in the multi-modal panoramic attention model described in this chapter is commonly used(Itti et al., 1998), therefore the other components of the model can be readily layered upon existing implementations of the same. An important benefit of the three-layer attention model is the ability of its mid-level layer to act as a spatiotemporal filter for sensory input. The panoramic map provides a high-level environmental assessment of entities around the robot that can be used to develop behavioral models or aid in task completion. The idle gaze behavior model combines exploratory gaze behavior, active gaze directed at new observations and subsequent gaze tracking behavior. Significantly, these benefits are obtained for free while providing the impression of liveliness. Finally, the

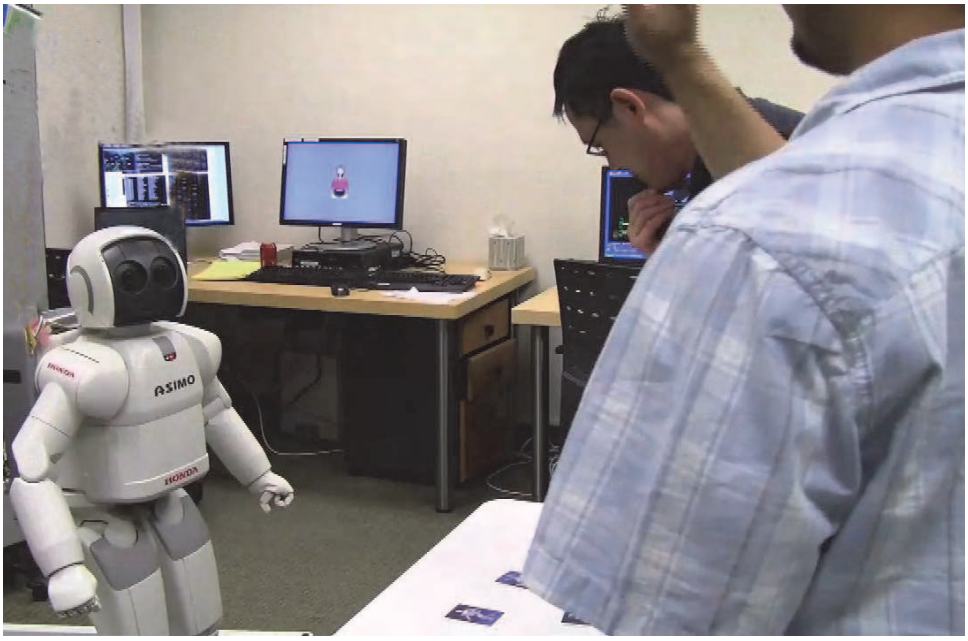


Fig. 7. Hand-waving to get robot's attention while it is playing card game with another person

incidental information obtained through casual observation can be quickly recalled when location-based queries are performed.

6.1 Limitations

The current attention pipeline can be accelerated significantly if the processing related to producing regions of interest is parallelized. This can either be done during low-level feature computation by utilizing GPU acceleration techniques or at higher levels by allowing regions of interest to be processed simultaneously by multiple types of detectors or the regions of interest could be paritioned spatially to be analyzed in parallel by the same type of detector. In a distributed system, the communication between the different levels of the attention model is handled using network socket communication. If there are numerous communication links from the raw sensory input to the highest panoramic attention layer, the cumulative latency will create a delayed response in perception and any subsequent behavioral reponse. By judiciously allow increased computational load on a single computer, network communication can be replaced by very rapid memory-based messaging schemes which will allow the robot to respond quicker to environmental changes.

To reduce excessive computational overhead, the use of context can be applied to adjust weighting of different features when determining regions of interest. For example, a highly visual task may not require frequent sound processing. The attention system can choose to either adjust the sampling rate of the sound detector or disable processing completely. Increased processing can then be allocated to relevant features to produce more accurate measurements and subsequent state estimates. At the middle layers, introducing a top-down mechanism of modulating different types of detectors can create efficient resource allocation.

Fig. 8. Automatic logging of speaker activity and locations for multi-speaker applications: (inset) Panoramic attention locates speakers and logs amount of speaker activity for each participant. White circle regions represent past clusters of sound activity labeled with current number of utterances and cumulative time spoken.

For example, if it is discovered that certain objects are not present in an environment or not relevant to a task, the corresponding detectors can be suspended to allow relevant detectors to perform at faster rates, improving overall system performance.
The mechanism for top-down modulation of the attention system should be handled by a behavior system that is driven by the set of active applications running on the robot. The behaviors are dictated by the current task and there should be a clear layer of separation between the behavior and perception system. Behaviors are triggered by changes of the environmental state which is inferred from the panoramic attention system. Therefore, it should be the behavior system that configures which detectors and low-level features the perception system should perform to allow proper decision-making for behaviors to be made. The behavior system can consist of either low-level reactive actions that may be triggered directly by low-level features, or high-level deliberative behaviors that may spawn multiple sub-tasks themselves. Since reactive behaviors have fewer intervening communication links, generated responses will automatically occur quicker in response to changes in sensory input. Since the locations in the panoramic map are in egocentric coordinates, they need to be updated whenever the robot moves to a new location. Although these locations can be completely re-sensed and re-calculated once the robot moves to its new position, the

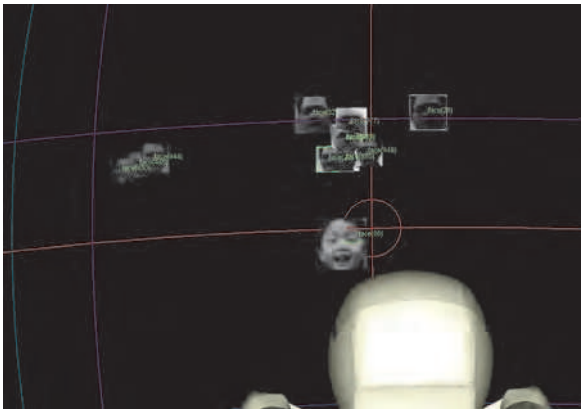


Fig. 9. Child faces appears lower in panorama.

panoramic map would be invalid during actual motion, limiting the robot's perception while it is moving. Continuously tracking the position of environmental objects during motion and incorporating the known ego-motion of the robot could be employed to create a continuously updated panoramic map even while the robot is moving.

6.2 Future work

To make the low-level attention model more consistent, it would be preferable to replace the skin and motion feature maps with their multi-resolution versions. By adding more cues to the low-level model such as depth information and audio source localization, the panoramic map can be extended to aid in navigation and manipulation tasks, as well as identification of speaker roles to follow conversations or identify the source of a query. This would provide a multi-modal basis to the existing model. If an object-recognition module is also integrated into the mid-level layer, it would widen the scope to scenarios involving objects and human-object interactions.

The semantic information stored in the panoramic attention map can be exploited to produce better motion models. For example, it is expected that a face belonging to a person is more likely to be mobile compared to an inanimate object resting on a table. Consequently, the dynamic motion models can be tuned to the entity type.

Spatial relationships in the panoramic map can also be used to deduce information about entities in the system. In Figure 9, one of the faces appears in a lower vertical position than the others. Given that the bounding boxes for both faces are in the same order of magnitude, the robot can infer that one person could be a child or sitting down. If the robot was not able to detect an inanimate chair in the region of the lower face, it could further deduce the face belongs to a child.

7. References

C. Breazeal & A. Edsinger & P. Fitzpatrick & B. Scassellati(2001). Active vision for sociable robots, In: *IEEE Transactions on Systems, Man, and Cybernetics, A*, vol. 31, pp 443-453.

L. Itti & C. Koch & E. Niebur(1998). A model of Saliency-based Visual Attention for Rapid Scene Analysis, In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp 1254-1259.

L.Itti & G. Rees & J.K. Tsotsos(2005). In: *Neurobiology of Attention*, Academic Press, 2005, ISBN 0123757312, 9780123757319.

L. Komatsu(1994). In: *Experimenting With the Mind: Readings in Cognitive Psychology*, Brooks-Cole Publishing Company.

L. Itti & N. Dhavale & F. Pighin(2003). Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention, In: *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*, vol. 5200, pp 64-78.

C.Kayser & C. I. Petkov & M. Lippert & N.K. Logothetis(2008). Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map, In: *Current Biology*, Vol. 15

P. Maragos(2008). In: *Multimodal Processing and Interaction: Audio, Video, Text*, pp. 181-184, Springer, ISBN 0387763155, 9780387763156.

K.Rapantzikos & G. Evangelopoulos & P.Maragos & Y. Avrithis(2007). An Audio-Visual Saliency Model for Movie Summarization, In: *IEEE 9th Workshop on Multimedia Signal Processing(MMSP)*, pp.320-323.

V. Navalpakkam and L. Itti(2005). Modeling the influence of task on attention, In:*Journal of Vision Research*, vol. 45, no. 2, pp. 205-231.

J.Moren & A.Ude & A.Koene(2008). Biologically based top-down attention modulation for humanoid interactions, In: *International Journal of Humanoid Robotics (IJHR)*, vol. 5, pp 3-24

B. Mutlu & T. Shiwa & T. Kanda & H. Ishiguro & N. Hagita(2009). Footing in Human-Robot Conversations: How Robots Might Shape Participants Roles Using Gaze Cues. In: *Proceedings of the 4th International Conference on Human-Robot Interaction (HRI)*.

Y. Nozawa & J. Mori & M. Ishizuka(2004). Humanoid Robot Presentation Controlled by Multimodal Presentation Markup Language MPML, In:*Proc. 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*.

T. Sobh & A. Yavari & H.R. Pourreza(2008). Visual Attention in Foveated Images, In:*Advances in Computer and Information Sciences and Engineering*, Springer, ISBN 9781402087417, pp. 17-20.

J. Shear & F.J. Varela(1999). In: *The View from Within: First-person Approaches to the Study of Consciousness*, Imprint Academic, ISBN 0907845258, 9780907845256

P. W. Halligan & D. T. Wade(2005). In: *The Effectiveness of Rehabilitation for Cognitive Deficits*, Oxford University Press, ISBN 0198526547, 9780198526544

A. Bur & A. Tapus & N. Ouerhani & R. Siegwar & H.Hiigli(2006), Robot Navigation by Panoramic Vision and Attention Guided Features, In: *International Conference on Pattern Recognition(ICPR)*, vol. 1, pp. 695-698.

K. Kayama & K. Nagashima & A. Konno & M.Inaba & H. Inoue(1998). Panoramic-environmental description as robots visual short-term memory, In:*IEEE International Conference on Robotics and Automation(ICRA)*, pp. 3253-3258.

K. Nickel and R.Stiefelhagen(2007). Fast Audio-Visual Multi-Person Tracking for a Humanoid Stereo Camera Head, In:*IEEE-RAS 7th International Conference on Humanoid Robots - Humanoids'07*, Pittsburgh, PA.

J. Ruesch & M. Lopes & A. Bernardino & J. Hörnstein & J. Santos-Victor & R. Pfeifer(2008), Multimodal saliency-based bottom-up attention framework for the humanoid robot iCub, In:*IEEE International Conference on Robotics and Automation (ICRA)*, pp. 962-967

A. Koene & J. Moren & V. Trifa & G. Cheng(2007). Gaze shift reflex in a humanoid active vision system, In:*Proceedings of the 5th International Conference on Computer Vision Systems (ICVS)*.

W. D. Stiehl & J. K. Lee & R. Toscano & C. Breazeal(2008). The Huggable: A Platform for Research in Robotic Companions for Elder care, In:*AAAI Fall Symposium, Technical Report FS-08-02*, The AAAI Press, Menlo Park, California.

J. Macedo & D. Kaber & M. Endsley & P. Powanusorn & S. Myung(1999). The effects of automated compensation for incongruent axes on teleoperator performance, In:*Human Factors*, vol 40, pp. 541-553.

S.M.Pizer & R.E. Johnston & J.P. Ericksen & B.C. Yankaskas & K.E. Muller(1990). Contrast-limited adaptive histogram equalization: speed and effectiveness, In:*Proceedings of the First Conference on Visualization in Biomedical Computing*, pp. 337-345.

N. J. B. McFarlane & C. P. Schofield(1995). Segmentation and tracking of piglets in images, In:*Journal of Machine Vision and Applications*, vol. 8, no. 3, pp. 187-193,Springer.

P. Kakumanua & S. Makrogiannisa & N. Bourbakis(2006). A survey of skin-color modeling and detection methods, In:*Pattern Recognition*, vol. 40(3), 2006

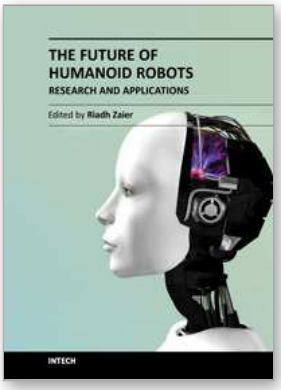
G. Bridgeman & D. Hendry & L. Start(1975). Failure to detect displacement of visual world during saccadic eye movements. In: *Vision Research*, 15, pp. 719-722.

I.D. Guedalia & M. London & M.Werman (1998). An On-Line Agglomerative Clustering Method for Nonstationary Data. In: *Neural Computation*, vol. 11(2), pp.521-540.

ASIMO humanoid robot (2000). Honda Motor Co. Ltd.

P.A.Viola & M.J.Jones (2004). Robust Real-Time Face Detection. In: *International Journal of Computer Vision*, vol. 57(2), pp.137-154.

S. Ravi Kiran & V. Ng-Thow-Hing & S.Okita, In: *The 19th IEEE International Symposium on Robot and Human Interactive Communication(RO-MAN)*. pp.7-14.



The Future of Humanoid Robots - Research and Applications

Edited by Dr. Riadh Zaier

ISBN 978-953-307-951-6
Hard cover, 300 pages
Publisher InTech
Published online 20, January, 2012
Published in print edition January, 2012

This book provides state of the art scientific and engineering research findings and developments in the field of humanoid robotics and its applications. It is expected that humanoids will change the way we interact with machines, and will have the ability to blend perfectly into an environment already designed for humans. The book contains chapters that aim to discover the future abilities of humanoid robots by presenting a variety of integrated research in various scientific and engineering fields, such as locomotion, perception, adaptive behavior, human-robot interaction, neuroscience and machine learning. The book is designed to be accessible and practical, with an emphasis on useful information to those working in the fields of robotics, cognitive science, artificial intelligence, computational methods and other fields of science directly or indirectly related to the development and usage of future humanoid robots. The editor of the book has extensive R&D experience, patents, and publications in the area of humanoid robotics, and his experience is reflected in editing the content of the book.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821