INTERACTIVE AUDITORY DISPLAY TO SUPPORT SITUATIONAL AWARENESS IN VIDEO SURVEILLANCE

Benjamin Höferlin[†], *Markus Höferlin*[‡], *Michael Raschke*[§], *Gunther Heidemann*[†], *Daniel Weiskopf*[‡]

[†]Intelligent Systems Group, [‡]VISUS, [§]VIS
Universität Stuttgart, Germany
{hoeferlin, hoeferms, raschkml, ais, weiskopf}@vis.uni-stuttgart.de

ABSTRACT

A key element for efficient video surveillance is situational awareness. Characteristics of human perception (e.g., inattentional blindness) as well as surveillance practice (e.g., CCTV operators have multiple responsibilities) often hinder comprehensive visual recognition of the activities in the monitored area. We support situational awareness and reduce the workload of CCTV operators by complementing the video display by an auditory display. Trajectories of moving objects extracted from surveillance video are sonified by auditory icons. These icons are interactively assigned by the user to each object category of the video and, in this way, form a sonic ecology. We use a spatial auditory display to represent location, direction and velocity of a trajectory with respect to a virtual listener. This facilitates orientation in virtual auditory space in a natural and realistic way that meets users' expectations. *Modification areas* are introduced to allow the users to define areas in which auditory icons are modified to further improve situational awareness. We put emphasis on efficient interaction between users and the auditory display to adjust the system according to the monitored area. Finally, we evaluate our approach by a user study and discuss benefits and shortcomings of the proposed sonification in the light of psychology, cognitive science, and neuroscience.

1. INTRODUCTION

The number of closed-circuit television (CCTV) cameras has rapidly increased within the last years. Due to the vast amount of cameras not all are monitored continuously and properly. Hence, detection of relevant events or threats cannot be guaranteed.

CCTV operators often have additional responsibilities that may distract their attention from active monitoring. These duties include the logging of incidents, communication with individuals inside and outside the control room, tape management, preparation of working copies for further investigation or evidence to the court, and controlling the entry/exit of the control room itself [1, 2]. Further, human needs, such as coffee breaks or the break for a smoke may interrupt continuous surveillance. Gill *et al.* observed control room operators being away from their screens in approximately 20% of their shift time [1].

Further, human recognition and visual perception are limited and may lead to missed events and undetected threats. After about 20 minutes of monitoring video screens attention of most individuals (even if dedicated and well-intentioned) will fall below an acceptable level [3]. This loss of attention is fostered by boredom due to little intellectually engaging stimuli in surveillance footage. Another shortcoming of human surveillance capability is *change blindness*. Change blindness describes the difficulties of the human perception to identify unexpected changes during blinks, flickers, or disruptions. Scott-Brown and Cronin point out that the average CCTV system can be considered as a change blindness machine due to flickers and interruptions while switching between cameras on screen [4] and when watching recorded time-lapse video with low temporal sampling resolution [5]. Gill and Spriggs [6] report that 9 of 13 evaluated control rooms record their video footage with less than 1.5 fps and Keval [2] observed that 8 of 14 evaluated control rooms use less than 8 fps, which is the minimum frame rate for effective crime detection [7]. While the issue with change blindness decreases if the object in change attracts attention [8], changes off a strong attentional focus are recognized poorly. This effect called *inattentional blindness* was impressively demonstrated by the "Gorillas in our Midst" experiment by Simons and Chabris [9]. In context of video surveillance, we may conclude that relevant events or threats might be missed, since operators focus on particular suspects[1].

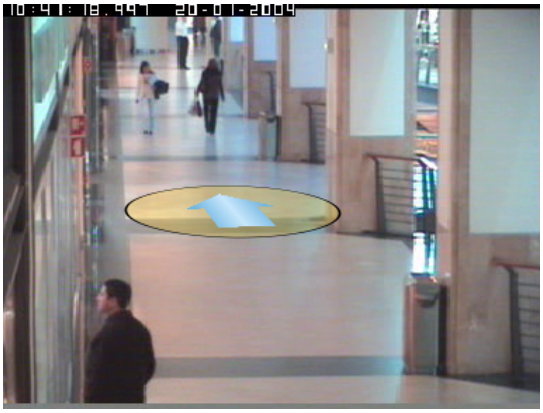

Figure 1: Interface of the auditory display superimposes a surveillance video from the CAVIAR dataset. Moving objects in video are sonified by auditory icons according to their object category (e.g., persons sound like steps).The blue arrow represents the virtual listener's position and direction in the spatial auditory display that is aligned to the floor in the video. The yellow circle (projected to the ground-plane) indicates the maximum distance of auditory icons that sound with maximum volume. Auditory icons outside the circle follow a logarithmic sound attenuation.

[1]According to Keval [2], operators look out for known offenders, previously observed crime patterns, and targets with revealing, unusual body language and negative emotions.

As a result, situation awareness of CCTV operators is reduced, and thus, their performance. In this paper, we will focus on the issue of losing attention due to additional duties and distraction from screen, as well as on the shortcomings of human recognition and perception. We alleviate these factors by supporting the human operators in their visual surveillance task by auditory cues. Particularly, we take advantage of the nature of surveillance video data that commonly lacks of audio tracks and is uncut, in contrast to narrative movies. Therefore, we introduce a novel spatial auditory display for video surveillance footage (see Figure 1). The proposed sonification combines spatial oriented auditory icons of moving objects with the parameter mapping approach, adapting the sonic properties of an icon according to the properties of its trajectory. Further, we introduce the concept of *modification areas* to improve interpretation of the auditory display. To achieve the goal of increased situational awareness and reduced workload we consider aspects of human cognition. Our second contribution is to define a user interface that facilitates interpretation of sound properties by providing context information in a way that meets users' expectations. Therefore, we introduce the concept of a *virtual listener* that could be placed anywhere in the video. We evaluate the introduced sonification approach by a user study and discuss its properties in the context of psychology, cognitive science, and neuroscience.

2. RELATED WORK

There is only limited coverage of sonification of video data in previous research literature. One example is the *Cambience* system by Diaz-Marino [10], which is able to map the video stream from webcams to a sonic ecology. The system uses difference images to estimate the activity in user-defined regions of the video stream. Features derived from change within the regions (e.g., center of change, velocity) are mapped to sound parameters, such as position in 2D audio space, volume, or playback frequency. Diaz-Marino considered Cambience to be useful in several application areas, such as interactive art, to provide informal awareness between collaborators, and as a security system that provides notification of change in video as audio information. Especially, the latter use case matches to some extent the scenario we cover in this paper. However, he rather emphasized visual programming of sonic ecology and social awareness systems than the support of human recognition in video surveillance tasks.

Pelletier [11] maps a sparse optical flow field to the parameters of its video sonification. This approach can be considered as tracking of interest points, such as corners. This trajectory sonification enabled him to express the performance of dancers in a musical way. The purpose of his approach is of creative and artistic nature, hence aesthetics of the sonification are in focus. Our approach of sonification of video surveillance data uses trajectory information, too. In contrast to the work of Pelletier, we utilize trajectories of moving objects, which are on a higher semantic feature level. Our sonification method is driven by cognitive and perceptual issues to improve performance in surveillance monitoring.

Other systems for trajectory sonification in literature mainly extract trajectories from one or multiple video streams, often utilizing motion capture systems, such as VICON [12, 13]. It is a common approach to use parameter mapping to connect trajectory features to sound parameters (e.g., [12]). Application areas for trajectory sonification primarily range from creative or artistic purposes [11] to the assistance of motor learning [12, 13] with application to sports or rehabilitation. In contrast to these approaches, we apply video sonification to the field of video surveillance. By combining visual and auditory display, we assist the users (typically CCTV operators) in their monitoring task. The goal of our sonification is to support their situational awareness and to reduce the workload.

3. PROBLEM DEFINITION

As we already mentioned in the introduction, situation awareness in video surveillance is often limited by the constraints of human recognition abilities. Situational awareness of CCTV operators depends on their distribution of visual attention. Strong focus on particular objects (see inattentional blindness [4]) is similarly bad as inattention due to distraction.

Situational awareness benefits from sonification of surveillance video data, since human sound perception is omnidirectional and ubiquitous. In contrast to vision, hearing does not require a particular direction of the listener's head-body configuration to perceive a desired signal. This allows the listener to move freely while hearing.

However, humans are not just aware of a situation because an acoustic signal reaches their ears. The question rather is, to which extent humans can handle multiple tasks and "split" their attention and processing resources among these. This question is addressed by the multiple-resource theory. Dual-task experiments indicated that structural dichotomies (e.g., such as visual and auditory processing) behave like separate resources [14]. This finding led to the 4-dimensional multiple resources model that claims increasing interference between two tasks to the extent that they share processing stages (perception/cognition, response), sensory modalities (auditory, visual), codes (verbal, spatial), and channels of visual information [15]. Hence, dual-task design can benefit from the use of separate resources [16].

This knowledge motivates us to use displays of different modalities to address the problem of dual-task interference. Concretely, we propose to use an auditory display besides conventional surveillance monitors. The goal of our interface design is to account for the concurrent tasks of CCTV operators (see [1, 2]) and to minimize interferences between the tasks. For example, users can monitor surveillance screens (visual/spatial) while simultaneously chatting with a colleague (auditory/verbal); or they can listen to the auditory display (auditory/spatial), while writing an incident report (visual/verbal, response). For a more comprehensive list of process-specific resources, we refer to the work of Boles [16]. The objective of auditory display is to attract attention whenever visual focus is away from relevant changes on the screen.

The human auditory system is very capable of recognizing changes in audio patterns. Various studies based on magnetoencephalography show that *mismatch negativity* (*MMN*, a change-specific component of the auditory event-related brain potential) elicitation is an inattentional automatic brain process. Furthermore, it is assumed that MMN initiates switching of attention to potentially important events in the unattended auditory environment. It was further shown that MMN generation is sensitive to various types of sound change, such as changes in frequency and temporal aspects (e.g., duration, gap in stimulus), but also shows tolerance in some range of the deviation to standard [17].

Since changes in auditory patterns can draw attention and guide focus to the object in change, one of the main goals of the proposed sonification is to map change in video proportional to

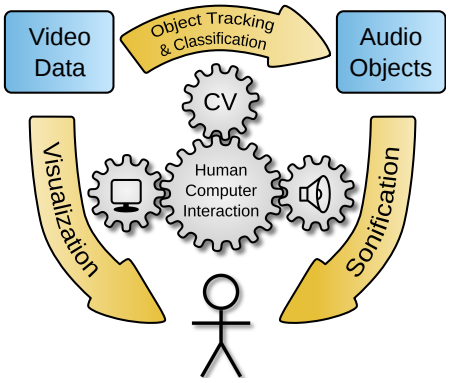

Figure 2: Structure of our sonification system. Trajectories of moving objects are extracted from video by computer vision and sonified as auditory icons by a spatial auditory display. The users gain situational awareness by monitoring audio and video signals, complementing each other. Users interact with the system by a graphical user interface to adapt parameters according to their task.

change in audio. This implies that small changes in video activity will result in small changes of the audio pattern. Thus, recurrent movement and activity can be perceived as a stationary sound pattern (also called an auditory texture). Pattern changes above a certain threshold may trigger the attention allocation mechanism of the auditory system.

According to the problem definition, our sonification approach includes components of multiple design patterns [18], such as *Ears Lead Eyes*, *Auditory Icon*, and *Situational Awareness*.

We summarize the requirements of an auditory display to support situational awareness in surveillance video:

- Auditory background: Usual activity should be perceived as auditory background pattern and should not attract attention.
- Allocation of attention: Variation from common activity as well as relevant actions should draw attention.
- Interpretation: Identification of objects and recognition of events should be facilitated.
- Multi-tasking: Dual (multiple) task performance should be improved and the experienced workload should be decreased.
- Adaptivity: Sonification has to be adaptable to different users and different areas under surveillance.
- Scalability: The auditory display has to cope with heterogeneous surveillance settings (perspective, number of objects) and different numbers of monitors and cameras.

4. SONIFICATION APPROACH

For sonification, we focus on relevant information of surveillance video footage to reduce the complexity of the auditory representation and to facilitate interpretation of the produced sounds. For this purpose, we use the trajectories of moving objects as they are relevant entities in video surveillance: static environments or objects generally do not represent any threat.

4.1. Extraction of Auditory Objects

Based on video data from surveillance cameras, we extract trajectories of moving objects using computer vision techniques. Therefore, we detect changing regions in video by applying the *ViBE* background subtraction [19] and track them by a linear Kalman filter [20]. Further, we classify the moving objects by the properties of their trajectories (e.g., movement speed and object size). Finally, the homography transform between the camera view and a virtual top view is calculated using a ground-plane assumption. We apply this transform to calculate position and speed of the tracked objects in real-world coordinates. Since computer vision is out of scope of this paper and the mentioned tracking pipeline is standard in automatic video analysis we will not discuss trajectory extraction in detail.An overview of the whole workflow is depicted in Figure 2.

4.2. Sonification of Auditory Objects

In contrast to the approaches of video/trajectory sonification in literature (see Section 2), we represent the data by a mixture of sonification techniques: auditory icon and parameter mapping, rather than a pure parameter mapping. In our approach, each trajectory is represented by an auditory icon according to its object category. Trajectories as well as their object category are provided by computer vision. For example, trajectories that belong to the category "people" can be acoustically represented by footsteps, and "cars" may sound like an engine. The advantage of natural sounds over earcons or artificial sounds is that users are familiar with the sounds and know how to interpret them, without the need to learn their interpretation. Hence, category information is conveyed in a natural way in which the sounds build a familiar sonic ecology.

Information about the position, direction, and velocity of objects' trajectories are fundamental in order to gain situational awareness in surveillance scenarios. The relative position of auditory objects is presented to the users by an 2D spatial auditory display. We only use two spatial dimensions, since we assume every relevant object to be located on the ground-plane that we used for tracking. Hence, auditory icons are only found on the horizontal plane. This additionally avoids more difficult elevation judgements of a sound source. Further, we apply parameter mapping to encode additional information of the trajectories, such as movement direction, and velocity, to the sound properties of the auditory icons. We map the properties of trajectories to the sound parameters in a way the users are familiar with. Therefore, we apply a physical model to the virtual 2D audio space that corresponds to real-world acoustics. To create such an auditory display, we utilize the FMOD sound system[2]. FMOD manages the virtual sound space and handles spatial sound distribution via level panning in surround settings (e.g., 7.1/5.1 speaker settings) or via a *head-related transfer function* (HRTF) in headphone settings. In detail, we use MyEars[3] calibration process to retrieve an approximated individual HRTF for 7.1 to stereo down-sampling, which is supported by FMOD. Please note that the spatial orientation of an auditory icon with respect to the properties of the moving object it represents, can be regarded as a special form of parameter mapping. By this means, the location of a moving object is mapped to interaural level difference, interaural time difference, pinna reflections, etc. In the same way, properties such as distance between

[2]FMOD Sound System, copyright (c) Firelight Technologies Pty, Ltd.
[3]http://www.myears.net.au/

object and listener are encoded by volume, sound roll-off, or early reverberation. The impression of the movement direction and the velocity of an auditory object is created by frame-wise update of its position, according to the extracted trajectory. Additionally, movement direction and velocity of a trajectory are mapped to the Doppler effect. To further amplify velocity perception, users can choose to map object velocity to playback speed of the auditory icon, which for example approximates the well-known effect of an accelerating engine. The benefit of parameter mapping being analogous to real-world perception is that object recognition and situational awareness are facilitated.

Please note that there are differences between the proposed sonification and potential surround/binaural recordings of the surveillance scene. An obvious benefit of our method is that the virtual listener (steered by the CCTV operator) is able to move to any position in the 2D auditory space; the listener is not fixed to the position of the recording device. A second advantage of video sonification over playback of natural audio recordings is the abstraction of the audio content. Inconvenient background noise, such as the blowing of the wind, is avoided in sonification. Further, our auditory display is rather schematic, which allows us to highlight relevant parts in auditory perception. For example, auditory icons of a person and a bus might be represented at the same volume and could both be recognized by the operators. In natural environments, the bus would drown out the person, without the operators being aware of the person. In the same way, auditory icons of trajectories could be accentuated following the schematic illustration of cartoons. Figures in cartoons are reduced to their relevant elements and main objects are disproportionately highlighted. The cartoon metaphor can be used to emphasize important properties of an auditory object and to neglect irrelevant details, too. Parameters that allow the users to put emphasis on particular object types, regions, or environmental settings, as well as the user interaction with the virtual listener are described in Section 4.4.

To get an impression of the capabilities of our sonification, we provide examples of our system, including an example that illustrates the user interaction experience, on our website[4]. Figure 1 depicts a frame of a sonified surveillance video superimposed by the representation of the virtual listener. For this example, we used video and ground-truth annotation of the CAVIAR dataset[5].

4.3. Modification Areas

Another concept we introduce with our sonification is the *modification area*. Modification areas are user-defined regions in the video context that affect the sound properties of auditory icons located in these regions in 2D auditory space. The concept of modification areas is derived from the observations of real-world effects. Their counterparts in real-world could be green areas aside the pavement or space enclosed by walls. For example, if pedestrians walk from the pavement across the grass and back to the pavement, the sound of their steps will change, even though the steps remain the same. For instance, the sound gets muted and high frequencies are cut off, while walking over grass. We observe an analogous alteration of sound between an area enclosed by walls and space without walls. In the first case, we expect more reverberation than in the latter.

Figure 3: Graphical user interface for the definition of a modification area. Left: specification of the region of influence by Photoshop-like brushing (the circle shows the brush). Right: specification of modification filters and their parameters.

The definition as well as the application of modification areas comprise two parts: the region in which a sound is affected by a modification and the modification of the sound properties. During video sonification, we check every frame if an auditory object entered or left a modification area. If such transition is detected the sound effects applied to its auditory icon are adapted. There are several filters and sound effects, such as lowpass, distortion, or complex reverberation settings, that can be applied by a modification area to an auditory icon. Further, the parameters of filters and sound effects can differ between different modification areas, resulting in a variety of combinations. Please note that a modification area does not replace the auditory icon of a trajectory, but only modifies its sonic properties. It is important to keep an object identifiable by its auditory icon, even when moving from one modification area to another. Obviously, modification areas can be used to convey properties of the environment, such as the material of a surface. However, usage of modification areas is not restricted to natural effects. In fact, they can be used in an abstract and more rich way, too. Users can, for example, define restricted areas or virtual trip-wires to get an auditory alarm, if people move into a sterile zone or a dangerous area. The definition of modification areas will be addressed in Section 4.4 in the context of user interaction.

4.4. User Interaction

Functional efficiency of video sonification (i.e., the capability to provide situational awareness and to reduce workload) depends largely on an appropriate selection of parameters for the auditory display. In turn, these parameters depend on the particular surveillance task, the monitored area, and further, on individual factors. Thus, it is essential to support close interaction between users and the system. In this context, interaction is not just limited to set some initial parameters, but is rather an explorative process in the virtual auditory space. Users can explore the monitored area by moving around the virtual listener to find a sweet spot that supports situational awareness (e.g., the center of a road junction to monitor car turning activity). Further, interaction allows the users to control the sound generation parameters to fit best their expectations. Besides optimal interpretation also aesthetics play a role in composing a non-obtrusive sonic ecology. It is important to allow the users to compose a non-obtrusive sonic ecology according to their

[4] http://www.vis.uni-stuttgart.de/projekte/visual-analytics-of-video-data/sonification.html. The audio signal is encoded in stereo channels with individually approximated MyEars HRTF, for use with headsets.

[5] EC Funded CAVIAR project/IST 2001 37540, found at URL: http://www.dai.ed.ac.uk/homes/rbf/CAVIAR/

preferences by defining auditory icons for each object category (e.g., person, car, etc.). We use multiple auditory icons for each object category to increase distinctness of different trajectories of the same class. The actual auditory icon will be selected randomly from the set of available icons using an uniform probability distribution. Alternatively, users can choose to apply a round-robin selection scheme to get most varying icons in the auditory scene. Capability of human perception to distinguish between multiple auditory icons is discussed in Section 5.2.

Presentation and manipulation of the properties of the auditory display in the context of video information is another important principle of the human-computer interface of our sonification approach. This facilitates interpretation of the sonification parameters. For this purpose, we augment the video by a representation of the virtual listener (see Figure 1). Besides information on position and viewing direction, the distances that define volume attenuation of audible sound objects are depicted, too. We use a projection of the virtual listener to the ground-plane of the video to further facilitate the understanding of the defined position, direction, and distances. Users can drag its representation to navigate the virtual listener to any position suitable for auditory surveillance. This interaction scheme encourages users to explore the auditory space and to find appropriate parameters for the auditory display in a way that meets their expectations. To receive an impression of the graphical user interface (GUI), we refer to the supplementary video[4]. Other properties of the auditory environment, such as the type of roll-off (linear/logarithmic) or the scale of Doppler shift can be adapted by a GUI. To also encourage the explorative usage of modification areas, their specification has to be simple. We provide a GUI that utilizes the brushing metaphor known from drawing applications, such as Photoshop. Users simply mark the extent of a modification area by brushing on an image captured from the video stream. Sound modifications for the selected area are then composed from a pre-defined enumeration of filters and effects. In Figure 3, the definition of a modification area is illustrated.

5. DISCUSSION

We now discuss to which extent our approach meets the demands for such a system as postulated in Section 3. We judge our system with respect to these requirements on the theoretical foundation of psychology, cognitive science, and neuroscience. Further, we provide a proof of concept study in which workload and situational awareness in a dual-task scenario was evaluated.

5.1. Auditory Background and Allocation of Attention

We discuss the capabilities of our sonification system to draw attention to relevant events together with its ability to attenuate background activity.

As already pointed out in Section 3, the human auditory system is an excellent change detector, sensitive to the variation of many sonic properties and deviation from abstract rules [17]. Pulvermüller and Shtyrov also suggested that lexical, semantic, and syntactic information of human speech is processed automatically without focal attention [21]. The proposed sonification maps most of the relevant changes in surveillance video to basic sound properties, such as frequency and spatial location. The filters used for modification areas also apply to such sound properties. Sound deviations in such properties are subject to automatic auditory change detection and orientation of attention indicated by MMN and P3a (an event-related potential often preceded by MMN) [22]. Note that this automatic orientation process is called preattentive, because it does not require attention [22]. Based on the evidence of early semantic processing of words [21], we hypothesize that also auditory icons can be preattentively analyzed for match with their semantic context. An example is the appearance of an auditory icon in untypical context (e.g., car on footpath). We believe the preattentive mechanism of the human auditory system is capable of reducing effects of visual inattentional blindness and distraction of visual focus.

The question if irrelevant activity in video is perceived as ambient auditory background or not, is closely related to preattentive change detection and allocation of attention. MMN is only elicited if sound deviation exceeds a particular threshold. Additionally, MMN requires a few preceding repetitions of a standard stimulus before being elicited at a deviating stimulus [17]. By reducing surveillance video to its relevant parts — trajectories of moving objects — we eliminate many sources of irritation, such as changes of brightness or dynamic background objects (e.g., waving trees). Hence, we introduce some kind of activity threshold that minimizes sound of insignificant change. Further, minor variations will be filtered out by the deviation threshold of the preattentive network. However, regularity of activity of the monitored area predominantly affects the allocation of attention. The development of a stationary sound pattern strongly depends on this regularity.

5.2. Interpretation of the Auditory Display

In proactive surveillance, there is often no specific search target or predefined event, CCTV operators have to look for. They search for something abnormal, an undefined threat. Hence, an auditory display cannot solely present audible alarms for particular events to the users. It rather has to provide a variety of information in a way that is interpretable. For this reason, our sonification approach conveys fundamental information on moving objects, such as their position and object category, by spatially aligned auditory icons. For such auditory display, separation and localization of the auditory objects is essential for an appropriate interpretation.

It was shown by Bronkhorst [23] that human localization performance only marginally differs between real sound sources and virtual sound sources created with individualized HRTFs. Hence, for our sonification approach localization of auditory objects is just limited by human auditory system. Further, our approach facilitates the separation of individual objects by utilizing well-known sounds (auditory icons) for object representation. To maximize segregation of icons, we additionally use a set of different auditory icons for each object category. This enables the users to distinguish between individual objects in scenes populated with multiple objects of the same type. In this context, separation is closely related to the identification of an auditory icon, which in turn requires training of the users to get familiar with the sounds. It was demonstrated in a study of Mäkelä *et al.* [24] that the identification of different "walking sounds" exhibits a strong learning effect. Without training, participants were able to identify 13% of the sounds correctly, after a short learning period identification increased to 66%. We expect a similar learning effect for the identification of other auditory icons, too. Further, the human auditory system principally is capable to separate multiple auditory channels, if the channels differ in one or more feature dimensions, such as pitch or orientation [22]. In their study on spatial audio displays for speech communication, Nelson *et al.* [25] provide evidence that spatial

separation of speech signals (as feature for the segregation of multiple sound streams) enhances the participant's ability to identify critical speech signals in competing message environments. In a similar way, spatial separation of auditory icons can support their separation and identification.

5.3. Multi-tasking

Besides general qualification of our auditory display for video surveillance, we have to question its effect on dual-task performance in the combination with a visual display. Therefore, we consider the prevalent tasks of a CCTV operator (see [1, 2]) and list their particular resource allocations according to Bole's enumeration of processing resources [16]. Note that only a rough categorization of the main resources involved in these complex tasks can be provided:

- Monitoring of screens (visual-spatial/spatial-attentive)
- Logging of incidents, communication via messenger/mail (visual-verbal/linguistic, manual)
- Tape management, preparation of copies (manual)
- Controlling the entry/exit of the control room (visual-spatial/spatial-attentive)
- Communication via phone (auditory-verbal/linguistic, vocal)
- Chatting with colleague (auditory-verbal/linguistic, visual-spatial/facial-figural, vocal)

According to the multiple resource theory (MRT), dual-task performance is inversely correlated with the degree of interference between the two tasks, by the means of shared modalities, mental processing resources, and response resources. Our first observation is that most tasks require only one modality, either the auditory or the visual. Hence, providing surveillance information for both channels principally allows the users to maintain situational awareness when involved in other duties besides monitoring. Furthermore, the dominance of the verbal code in resource allocation of additional duties supports our decision to apply sonification to transform video data into auditory information. According to the definition by Kramer *et al.*, sonification "is the use of nonspeech audio to convey information" [26]. Based on MRT, we suggest that our non-speech auditory display only exhibits little interferences with any of the enumerated tasks that utilize the visual modality. Even in chatting situations where both modalities can be involved, acceptable dual-task performance is expected, because the proposed sonification utilizes non-verbal, auditory-spatial processing resources.

5.4. Adaptivity

Our sonification allows the user to interactively adapt its parameters to different surveillance scenarios and to a variety of user preferences. In video surveillance, video data obviously represents the main data source and the visual modality is the familiar way of monitoring. Thus, we use video as context for audio parameters, where it is applicable. This facilitates an easy understanding of the parameters and meets users' expectations. As described in Section 4.4, users place the virtual listener in the video and exploratively adapt its parameters having direct visual feedback. Furthermore, the GUI allows the user to compose and adapt a sonic ecology of the site under surveillance.

5.5. Scalability

Scalability by the means of adaption to different users and to different monitored sites is covered by the aforementioned adaptivity of our system. However, the question how multiple video feeds could be integrated in our auditory display is still open.

Although the expansion of our sonification approach to multiple video streams is out of scope of this paper, we will offer some possible directions for this problem, since it is intrinsically tied to our application.

- The trivial approach is to auditorily display only one camera at a time, with either automatically or user-defined switching between the cameras.
- In scenarios where multiple cameras monitor the same area, objects of all cameras can be aggregated into a common coordinate system, with respect to their geographical locations.
- In contrast to aggregation with respect to the cameras' geographical locations, surveillance videos can be integrated into a single auditory context according to the screen layout of the control room. This aggregation scheme provides auditory location cues to the screen where the activity is displayed.
- Another approach is to superimpose multiple independent auditory displays. To segregate the sounds of different displays, sound features can be applied that are not used in the actual approach, such as the vertical orientation of the sound source. The study of Veltman *et al.* [27] on pilots' task performance in fighter cockpits with 3D audio support indicates that information of two independent spatial auditory displays can principally be processed.

6. PROOF OF CONCEPT USER STUDY

The goal of the user study was to evaluate the effect of sonification support in video surveillance tasks. We evaluated the situational awareness and the workload of users in a dual-task scenario with and without support of our auditory display. The proof of concept was designed as a within-subjects user study.

Experimental Setup. The user experiment was conducted in a laboratory that was insulated from outside distractions. Two videos were presented on a 17 inch EIZO TFT screen at a resolution of 1280×1024 pixels with 24 bit color depth. The video stream was presented with 768×576 pixels and 25 fps with the VLC media player.

Stimuli and Tasks. The videos for the user study originated from the corridor view of the CAVIAR dataset. Since the dataset provides several videos of short durations between 0:15 min and 2:28 min, we concatenated these videos and marked the transitions by an acoustic signal (for the sonified video as well for the original video). The resulting two videos had a duration of about 10 min. The position and viewing direction of the virtual listener in the sonified version of both videos was set similar to Figure 1. For the sonification of the people, three different auditory icons of steps were used. The playback speed adaption of the icons according to the velocity of the objects was activated. We evaluated the workload and the situational awareness of the subjects when performing the task:

T: Count persons that leave the corridor through a specific door.

In parallel, the subjects had to highlight verbs in a given text as distractor task.

Subjects. Sixteen participants (average age 27 years, minimum 19 years, maximum 38 years). Sex was not considered as confounding factor for this study. All participants were students of our university. Thirteen subjects were computer scientists, one studied technology management, one economy and one linguistics. Subjects were paid € 10 for participation. All participants had normal or corrected-to-normal vision. An audiometry showed that all participants had normal hearing. Eight of the participants played an instrument or were members of a choir.

Study Procedure. First, subjects had to fill in a questionnaire about their age, field of study or profession, computer skills and grammar knowledge. They could state whether they played an instrument or were members of a choir. Then, they read a three-page instruction manual for both the sonified and not sonified surveillance task and the parallel task. After the participants were given time to read this tutorial, we did a practice run-through of the tasks. The duration of the complete training was 15 minutes. During this practice test, subjects could ask questions about both tasks and clarify potential problems or misinterpretations. We also used the practice test to confirm that the subjects understood both the surveillance task and the parallel task.

Then, we continued with the main evaluation that took 25 minutes. Subjects had to perform the main task and the parallel task simultaneously, once with a sonified and once with a not sonified video. We counter-balanced video and text for both parts to compensate for learning effects. There was a "Give Up" option, but it was not used by the subjects. To compare situational awareness of both parts, we measured the accuracy rates in performing the task. Additionally, we video-recorded the subjects during the study and analyzed the subjects' point of focus shifts between screen and text. After both parts, subjects had to fill out the NASA Task Load Index questionnaire (NASA-TLX). A second questionnaire was given to the subjects after the main evaluation in which they marked their preferences in using one of the two surveillance techniques. Finally, participants were given the opportunity to provide open, unconstrained comments.

Study Results. To compare both techniques we measured the accuracy rates in performing the main task. Subjects identified 14 people (in median) leaving the corridor in case of the sonified video stream (see Figure 4 (right))). When using the not sonified video stream they identified 15 people in median. The correct answer was 23 people in both cases. A Wilcoxon signed-rank test showed no significant differences between the two parts. Additional, t-test showed no significant differences in focus shift frequencies for both parts.

Figure 4 (left) shows the results of the NASA-TLX. This questionnaire asked: 1.) How mentally demanding was the task? 2.) How physically demanding was the task? 3.) How hurried or rushed was the pace of the task? 4.) How successful were you in accomplishing what you were asked to do? 5.) How hard did you have to work to accomplish your level of performance? 6.) How insecure, discouraged, irritated, stressed, and annoyed were you? Answers could be given in Likert-scale: Question 1-3 and 5,6 with 0 = very low, 20 = very high; Question 4 with 0 = perfect, 20 = failure. In the second questionnaire, subjects stated with 2.4 (Likert-scale from 1 = I agree to 5 = I disagree), that sonification

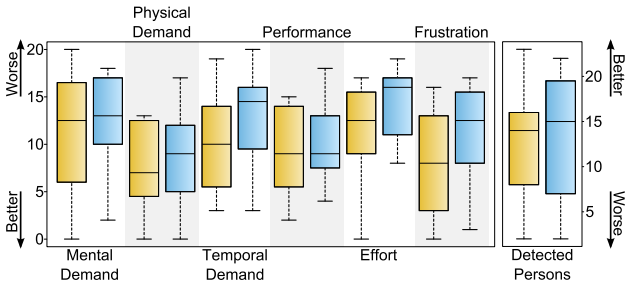

Figure 4: Boxplot of the user study results. Video display and auditory display in yellow; video only in blue. On the left: the results of the NASA-TLX. Answers could be given in Likert-scale (see main text). On the right: accuracy rates of the main task.

was helpful for them (standard deviation: 1.0), with 2.6 that they enjoyed sonification (standard deviation: 1.0) and with 2.6 that sonification by step sounds was helpful (standard deviation: 1.0). With the last question, we asked for the subjects' preferences in using one of the two techniques for a surveillance task. Fourteen subjects would prefer sonification, two the not sonified video.

Discussion. The comparison of the correctly identified persons shows that neither the identification quality nor the number of subjects' focus switches between monitor and text show significant differences between the sonified or not sonified videos. The results of the NASA-TLX in Figure 4 shows that in case of the sonified video the subjective workload is lower. According to Veltman *et al.* [27] this is a typical result, since a positive effect often exhibits in either performance or in workload, as they identified for the introduction of 3D auditory displays. Subjects were less hurried when performing the parallel task when using a sonified video. They also felt more successful with the sonification technique and had to work harder to accomplish the level of their performance with the not sonified technique. To conclude, the sonification technique leads to a less stressful and more comfortable surveillance.

7. CONCLUSION AND FUTURE WORK

In this work, we have introduced a 2D spatial auditory display to support situational awareness in video surveillance. We have applied a mixture of auditory icons and parameter mapping to sonify trajectories of moving objects extracted from video data. Besides appropriate mapping of object properties to sound properties that meets users' expectations and is well supported by human perception, we have put emphasis on an explorative user interface to facilitate a close feedback loop between users and auditory display. Finally, we validated the proposed sonification against recent results of research in psychology, cognitive science, and neuroscience. The proof of concept user study showed that the detection rate for both techniques is approximately equal. However, sonification leads to a lower workload and thus to less stressful and more comfortable surveillance.

Future research directions will include the sonification of more complex entities, such as human action (e.g., walk, run, box), interaction between moving objects (e.g., chatting, fighting, opening the car's door), and groups of objects (e.g., similar auditory icons for people that walk in a group).

8. ACKNOWLEDGEMENTS

The authors thank Tanja Blaschke and Manisha Singh for evaluating the user study and Michael Wörner for voice acting. This work was funded by German Research Foundation (DFG) as part of the Priority Program "Scalable Visual Analytics" (SPP 1335).

9. REFERENCES

[1] M. Gill, A. Spriggs, J. Allen, M. Hemming, P. Jessiman, D. Kara, J. Kilworth, R. Little, and D. Swain. (2005, 2) Control room operation: findings from control room observations. On-line Research, Development and Statistics publication. Home Office, UK. [Online]. Available: http://homeoffice.gov.uk/rds/pdfs05/rdsolr1405.pdf

[2] H. Keval, "Effective, design, configuration, and use of digital cctv," Ph.D. dissertation, University College London, 2009.

[3] M. Green, J. Reno, R. Fisher, L. Robinson, A. General, N. Brennan, D. General, J. Travis, R. Downs, and B. Modzeleski, "The appropriate and effective use of security technologies in us schools: A guide for schools and law enforcement agencies series: Research report," National Institute of Justice, Tech. Rep., 9 1999.

[4] K. Scott-Brown and P. Cronin, "Detect the unexpected: a science for surveillance," *Policing: An International Journal of Police Strategies & Management*, vol. 31, no. 3, pp. 395–414, 2008.

[5] ——, "An instinct for detection: Psychological perspectives on CCTV surveillance," *The Police Journal*, vol. 80, no. 4, pp. 287–305, 2007.

[6] M. Gill and A. Spriggs, *Assessing the impact of CCTV*. Home Office Research, Development and Statistics Directorate, 2005, Home Office Research Study 292.

[7] H. Keval and M. A. Sasse, "To catch a thief – you need at least 8 frames per second: The impact of frame rates on user performance in a CCTV detection task," in *Proceedings of ACM International Conference on Multimedia*. ACM, 2008, pp. 941–944.

[8] R. Rensink, J. O'Regan, and J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological Science*, vol. 8, no. 5, pp. 368–373, 1997.

[9] D. Simons and C. Chabris, "Gorillas in our midst: sustained inattentional blindness for dynamic events," *Perception*, vol. 28, pp. 1059–1074, 1999.

[10] R. Diaz-Marino, "A visual programming language for live video sonification," Master's thesis, University of Calgary, 2008.

[11] J. Pelletier, "Sonified motion flow fields as a means of musical expression," in *Proceedings of the 2008 International Conference on New Interfaces For Musical Expression*, 2008, pp. 158–163.

[12] A. Kapur, G. Tzanetakis, N. Virji-Babul, G. Wang, and P. R. Cook, "A framework for sonification of VICON motion capture data," in *Conference on Digital Audio Effects*, 2005, pp. 47–52.

[13] A. Effenberg, J. Melzer, A. Weber, and A. Zinke, "Motionlab sonify: A framework for the sonification of human motion data," in *Proceedings of Information Visualisation*. Computer Society Press, 2005, pp. 17–23.

[14] C. Wickens, "The structure of attentional resources," in *In R. S. Nickerson (Ed.), Attention and Performance VIII*. Lawrence Erlbaum Associates, 1980, pp. 239–257.

[15] ——, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, no. 2, pp. 159–177, 2002.

[16] D. Boles, "Multiple resources," *International Encyclopedia of Ergonomics and Human Factors*, pp. 271–275, 2001, in: Ed. Waldemar Karwowski; Taylor and Francis, London.

[17] R. Näätänen, P. Paavilainen, T. Rinne, and K. Alho, "The mismatch negativity (MMN) in basic research of central auditory processing: A review," *Clinical Neurophysiology*, vol. 118, no. 12, pp. 2544–2590, 2007.

[18] S. Barrass, "Sonification design patterns," in *Proceedings of the International Conference on Auditory Display*, 2003, pp. 170–175.

[19] O. Barnich and M. Van Droogenbroeck, "ViBE: A powerful random technique to estimate the background in video sequences," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE Signal Processing Society, 2009, pp. 945–948.

[20] G. Welch and G. Bishop, "An introduction to the Kalman Filter," Department of Computer Science, University of North Carolina at Chapel Hill, Tech. Rep. TR 95-041, 2004.

[21] F. Pulvermüller and Y. Shtyrov, "Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes," *Progress in Neurobiology*, vol. 79, no. 1, pp. 49–71, 2006.

[22] A. Johnson and R. Proctor, *Attention: Theory and Practice*. Sage Publications, Inc, 2004.

[23] A. Bronkhorst, "Localization of real and virtual sound sources," *The Journal of the Acoustical Society of America*, vol. 98, pp. 2542–2553, 1995.

[24] K. Mäkelä, J. Hakulinen, and M. Turunen, "The use of walking sounds in supporting awareness," in *Proceedings of the International Conference on Auditory Display*, 2003, pp. 6–9.

[25] W. Nelson, R. Bolia, M. Ericson, and R. McKinley, "Spatial audio displays for speech communications: A comparison of free field and virtual acoustic environments," in *Human Factors and Ergonomics Society Annual Meeting Proceedings*, vol. 43, no. 22. Human Factors and Ergonomics Society, 1999, pp. 1202–1205.

[26] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, J. Neuhoff, R. Bargar, S. Barrass, J. Berger, G. Evreinov, W. T. Fitch, M. Gröhn, S. Handel, H. Kaper, H. Levkowitz, S. Lodha, B. Shinn-Cunningham, M. Simoni, and S. Tipei, "Sonification report: Status of the field and research agenda," International Community for Auditory Display, Tech. Rep., 1999.

[27] J. Veltman, A. Oving, and A. Bronkhorst, "3-d audio in the fighter cockpit improves task performance," *International Journal of Aviation Psychology*, vol. 14, no. 3, pp. 239–256, 2004.