AUDITORY SUPPORT FOR SITUATION AWARENESS IN VIDEO SURVEILLANCE

Benjamin Höferlin[†], *Markus Höferlin*[*], *Boris Goloubets*[†], *Gunther Heidemann*[‡], *Daniel Weiskopf*[*]

[†]Institute for Visualization and Interactive Systems, Universität Stuttgart, Germany
[*]Visualization Research Center, Universität Stuttgart, Germany
[‡]Computer Vision Group, Institute of Cognitive Science, University of Osnabrück, Germany
hoeferlin@vis.uni-stuttgart.de

ABSTRACT

We introduce a parameter mapping sonification to support situational awareness of surveillance operators during their task of monitoring video data. The presented auditory display produces a continuous ambient soundscape reflecting the changes in video data. For this purpose, we use low-level computer vision techniques, such as optical-flow extraction and background subtraction, and rely on the capabilities of the human auditory system for high-level recognition. Special focus is put on the mapping between video features and sound parameters. We optimize this mapping to provide a good interpretability of the sound pattern, as well as an aesthetic non-obtrusive sonification: precision of the conveyed information, psychoacoustic capabilities of the auditory system, and aesthetical guidelines of sound design are considered by optimally balancing the mapping parameters using gradient descent. A user study evaluates the capabilities and limitations of the presented sonification, as well as its applicability to supporting situational awareness in surveillance scenarios.

1. INTRODUCTION

The goal of video surveillance is to spot irregular, abnormal, or suspicious behavior of persons and objects to identify and prevent illegal or threatening actions. The huge increase of closed circuit television (CCTV) installations over the last decade shows that video surveillance has been recognized to be an appropriate method for crime prevention and evidence recording. Though, in contrast to the rapidly growing number of surveillance cameras, the monitoring capabilities stay far behind this development. The reasons are manifold, but a major factor is the high expense associated with human resources. The extent of the imbalance between recording and monitoring capabilities becomes obvious in the high camera-to-operator ratio. In their observation of 13 control rooms, Gill *et al.* [1] came across camera-to-operator ratios from 20:1 to 520:1. Keval [2] reports camera-to-operator ratios from 4:3 to 120:1 in his study of 14 control rooms. In addition to the large number of cameras to monitor, operators are often responsible for a wide variety of other tasks. A brief enumeration of such additional and often concurrently processed tasks includes [1, 2]:

- logging of incidents,
- preparation of working copies for evidence to the court or further investigation,
- tape management,
- communication with individuals inside and outside the control room, and

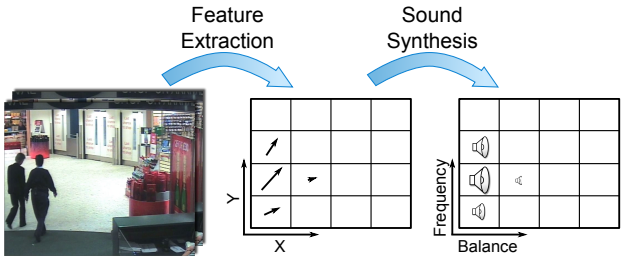


Figure 1: Segment-based feature processing and mapping to auditory parameters.

- controlling the entry/exit of the control room.

Such responsibilities lead to distraction from monitoring and hinder the detection of relevant actions and events. Further, human perception is subject to limitations that constrain the operator's event recognition ability. Such perceptual characteristics that have a strong influence on video surveillance performance include:

- the short *period of attention* when monitoring video screens (approximately 20 minutes [3]),
- difficulties to identify unexpected changes during blinks, flickers, or disruptions, called *change blindness* [4], and
- poor recognition of changes that are outside the focus of attention, termed *inattentional blindness* [5].

All these issues (mismatch of camera-to-operator ratio, additional responsibilities of CCTV operators, and perceptual constraints) point out that acceptable task performance in such high stress, multiple tasks environment requires proper situational awareness of the operators. As demonstrated by Höferlin *et al.* [6], sonification of surveillance data can support situational awareness and reduce subjective workload in multiple task scenarios.

In this paper, we apply feature extraction from video and map these features to auditory parameters (cf. Figure 1). One advantage of applying sonification to video surveillance is the complementary modality of the auditory display to the visual display, which is especially helpful when multiple target tracking and recognition tasks are performed [7]. According to the multiple resource theory, only a small degree of interferences of cognitive resources is expected in dual-task scenarios that require different mental modalities [8]. Such dual-task scenarios are typical in video surveillance [6]. Situational awareness in video surveillance further benefits from the complementary auditory display due to the excellent ability of the human auditory system to detect small changes in

sound patterns and to attract attention to those changes. As various studies pointed out (for a comprehensive overview see [9]), human auditory recognition is able to mask specific (e.g., recurrent) sound patterns from attentional processing, while being still sensitive to small variations of the sonic properties as well as to deviations to abstract rules, such as lexical, semantic, and syntactic information of human speech [10]. Such preattentive detection of change is often followed by orientation of the auditory focus of attention to the source (or auditory channel) of change. Preattentive change detection and subsequent switching of attention was well explored by magnetoencephalographical studies that explain these phenomena by differences in change-specific components of the auditory event-related brain potential, such as the *mismatch negativity* (MMN) [11].

Our approach exploits these beneficial properties of human auditory processing to support situational awareness in video surveillance. A basic assumption we make is that information relevant to surveillance monitoring is represented by changes in video signal. This means that we ascribe static parts of the video little or no relevant information. To leverage change detection capabilities of the human auditory system, our approach produces a continuous sonic pattern or soundscape of the change in video data. Further, recurrent changes in video generate an auditory texture that fades from attentional monitoring after some time of familiarization. In this state of background monitoring, sufficiently large changes of the auditory texture with respect to the familiar acoustic reference pattern reallocate attention, again. This is supported by research of the central auditory processing system that proved that MMN is only elicited after a few repetitions of a standard stimulus and only if the deviation exceeds a particular threshold [9]. Hence, we focus on the design of a non-obtrusive auditory display. Further, the parameter mapping should, to some extent, allow the interpretation of the sonification to infer from auditory display some information of the event that occurred in the monitored video. This supports a rough classification of the change recognized by the auditory signal and thus enables decision making, such as if the occurred event requires further attention by switching the visual focus to a screen.

These two main criteria for the design of our auditory display (interpretability and non-obtrusiveness) are reflected by the emphasis of this paper: the optimization between aesthetical and psychoacoustic aspects of this sonification. The goal is to find an aesthetically pleasing sonification that still conveys all of the relevant information in an interpretable manner.

1.1. Related Work

Little work has yet been published in the field of video sonification. Moreover, most of these sonifications were developed for artistic purposes (e.g., [12]) or as assistance of visually impaired people (e.g., [13]). In the context of video monitoring, we identified two related publications.

The first one is the *Cambience* system, which was developed by Diaz-Marino [14]. Besides its application in interactive arts, and as a technique to provide informal awareness between collaborators, Cambience was intended by its developer to be used as a security system that provides auditory alarms or notifications when changes occur in video. Therefore, Cambience maps video data from webcams to a sonic ecology. Differences between video frames are used to measure the level of activity in a video. Features derived from the level of activity in user-defined regions (e.g., amount of change, center of activity, and velocity) are mapped onto sound properties, such as volume, playback frequency, and stereo panning. Visual programming allows interactive definition of the mapping between sounds parameters and the features extracted from areas of interest. In the security context, Cambience provides an auditory display for process monitoring. This is closely related to the scenario we present in this paper. However, there is a distinct difference in the complexity of activities that are monitored between Cambience and the sonification presented in this paper. Cambience relies on user-defined areas of interest and is fixed on the recognition of apriori known events, such as a person entering a room. For this reason, it is constrained to be used mainly for auditory alarms. Abnormal behavior and more complex actions are thus hardly recognizable. In contrast, our approach is designed to guide attention also for apriori unknown activities and complex events that occur in the context of video surveillance.

The system by Höferlin *et al.* [6] utilizes trajectories of moving objects extracted from video data to support situational awareness of surveillance operators via a spatial auditory display. In their approach, each object trajectory is mapped to an auditory icon that moves along the object's trace in 3D sound space. By user interaction, the virtual listener's position and other parameters can be adopted to suit the monitored site. Further, the selection of auditory icons for each object class help produce a natural sound environment. The approach presented in this paper, follows a different path: one of the major differences is that we do not rely on high-level computer vision techniques, such as object tracking and classification, since these methods come with high computational cost and are not fully reliable [15]. Another difference is that we intend to avoid the mental reconstruction of the video from the auditory display. Such a translation from auditory stimulus to familiar mental representation was observed many times [16]. However, in the case of video sonification, maintenance of an imaginary video representation can be mentally demanding. We aim for a rather abstract auditory representation of relevant information and rely on the excellent capabilities of human auditory perception to detect deviations in the acoustic pattern. Although we aim for interpretability of the sonification, our primary goal is to enable auditory change detection on signal level, not on semantic level.

1.2. Contribution

According to the problem definition and related work, we aim for an auditory display meeting the following requirements:

- usage of reliable low-level computer vision features,
- comprehensive and abstract auditory display to leverage auditory change detection on signal level,
- synthesis of non-obtrusive continuous soundscape, and
- interpretability of the sonification to guide visual attention.

In the remainder of the paper, we present a novel parameter mapping sonification that copes with these requirements. This is our main contribution. As a major aspect, we tackle the often discussed issue of finding a trade-off between interpretability and aesthetics of sonification using non-linear optimization. Further, we evaluate our sonification with respect to its interpretability and support of situational awareness in video surveillance.

2. SONIFICATION DESIGN

To support the situational awareness in video surveillance, we propose a sonification system with the structure outlined in Figure 2. Besides the video display, users are provided with an auditory display based on low-level features extracted from video. These features are subsequently mapped to sonic properties of the continuous sonification. Our research prototype uses the CSound toolkit[1] for offline sound synthesis. Besides adjustment of a small set of parameters to select precision and mapping range, the auditory display does not need user intervention. Adapted values are not directly applied to the sonification, but used as input for parameter optimization to find an appropriate mapping with respect to aesthetic and psychoacoustic constraints of the auditory system.

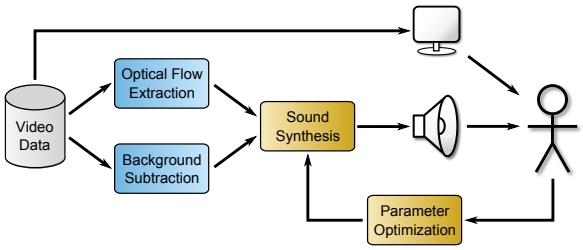


Figure 2: Data flow of the auditory display. Blue boxes depict data preparation steps by computer vision techniques. The yellow boxes represent the steps necessary for the parameter mapping sonification, described in this paper.

2.1. Data Preparation

Since we assume that only changes in video data are relevant for surveillance monitoring, we use as basic feature the dense optical flow field of two subsequent video frames. We extract the optical flow using the global method of Horn and Schunck [17]. The advantage of extracting dense optical flow over fast to compute frame differences is the availability of size and velocity information of the moving objects. For frame differencing this information is not available in the case of homogeneous colored objects, whereas the global optimization method of Horn and Schunck fills in the missing flow information by a regularization term. In addition to the motion vectors, we calculate a running average background model for foreground segmentation of the video data. This step is necessary, since optical flow calculation is prone to errors in the presence of noise and coding artifacts. Hence, motion vectors calculated in background regions are neglected for further processing. This approach helps reduce background noise and thus decrease obtrusiveness of the auditory display.

Next, we split the optical flow field into non-overlapping segments aligned in a regular grid as illustrated in Figure 1. For each segment, we calculate the average length of the contained motion vectors. This value represents the extent of activity for each segment. Please note that both the number of moving pixels and the length of the motion vectors (i.e., the velocity) influence the activity value. Hence, there are three properties for each segment to be mapped to auditory parameters: the segment's horizontal coordinate, its vertical coordinate, and its activity.

2.2. Mapping Function

There are many possible design choices for mapping the segment properties to sound parameters. However, preliminary experiments considering the users' expectations suggest the use of:

- stereo panning representing horizontal position component,
- frequency to represent the vertical component of position (rising frequency with increasing position), and
- amplitude to represent activity (low activity - soft sound).

Stereo panning and frequency dimensions are quantized, whereas amplitude is a continuous parameter. The directional information of motion in the segments is neglected. However, the direction of object movement is indirectly encoded in the temporal transition of the amplitude level between neighboring segments.

[1]CSound homepage: `http://www.csounds.com/`

From another point of view, each segment can be regarded to play its own instrument that is defined by stereo panning and frequency. If a segment shows no activity, the according instrument is muted. The complete orchestra of instruments represents the auditory display. Without aggregation to segments, motion features would be too sensitive to noise, or features of higher processing levels (e.g., trajectories) have to be used, which are prone to errors. Segmentation allows efficient sonification of low-level feature.

A key requirement of the auditory display is to convey the relevant information in an interpretable fashion. Additionally, the sonification has to be aesthetically pleasing to be non-obtrusive and broadly accepted [16]. To achieve these goals, we account for psychoacoustic aspects when defining the mapping and transfer functions. A formative user study (see Section 3) emphasized the importance of psychoacoustic aspects.

Pure tones are perceived to be unnatural, thus we use complex tones to increase natural sound sensation. For sound synthesis, each segment is represented by a periodic waveform synthesized by an additive synthesis model with 8 harmonics. Hence, the number of harmonic components we consider in the experiments is $N_H = 8$. Please note that we add only overtones that are whole multiples of the fundamental frequency in order to maintain pitch perception of complex sounds. Users can adjust the numbers of harmonics, if desired. However, although natural sounds generally have an arbitrary number of harmonics, their amplitude drops fast with higher harmonics. Thus, only few are audible and necessary for an almost natural sound sensation. By using a sine wave generator instead of MIDI sonification, we are able to tune the perceptual parameters of the sonification in much more detail, as described below. Employing the orchestral metaphor again, data features of each segment are mapped to perceptually calibrated *mini instruments* as proposed by Grond and Berger [18]. By adjusting the number of segments in each direction (horizontal and vertical), the users can trade the resolution and precision of the sonified information for the complexity of the produced soundscape.

The temporal sampling rate of the continuous sonification is set to the temporal resolution of the video data, and phases of the sine waves are adapted according to this rate to produce the impression of a continuous signal. We assume that the temporal sampling of online surveillance footage ranges from 15 fps to 30 fps. Hence, the temporal resolution of the human auditory system is capable of detecting sound changes between two successive frames. Typically, the temporal resolution for auditory change detection is beyond 20 ms, even for low frequencies (cf. [19]).

Further, we describe how we selected the transfer functions for each mapped parameter. To consider aesthetics and interpretability, we map the data properties not directly to physical sonic properties, but introduce an intermediate perceptual mapping layer.

2.3. Amplitude Mapping

To achieve linear scaling of amplitude that is necessary to interpret the information conveyed by the auditory display in the right way, we linearly map the activity value of a segment to the perceptual measure of subjective loudness S (sone at 1 kHz). Thereby, we scale the activity level to the sone interval that fits into the user-defined volume range. For the evaluation in Section 3, this range is fixed to the interval of 20 to 80 dB in the accordingly defined interval of frequency. Next step is to map loudness S to loudness level L (phon at 1 kHz) according to the non-linear relation [20]:

$$L = \begin{cases} 40 + 10\,\mathrm{ld}(S), & \text{if } S > 1 \\ 40S^{0.379}, & \text{else} \end{cases} \quad (1)$$

Finally, we map the loudness level with respect to equal-loudness-level contours to sound pressure level (dB-SPL); this value is directly fed into the CSound system and represents the amplitude of the fundamental frequency. Amplitudes of overtones are adapted accordingly and normalized by CSound. An analytical expression of equal-loudness-level contours fitted to experimental data was developed by Suzuki and Takeshima [21].

Obviously, this approach is only a rough approximation to adjust the perceived loudness of a data segment. We neglect any influence of overtones of complex sounds. Furthermore, dependencies between the complex tones of different data segments are not considered, too. A more elaborated loudness model will be considered in future work, a thorough evaluation of advanced models was presented by Skovenborg and Nielsen [22].

2.4. Stereo Panning

A segment's horizontal position component is a linearly mapped between left and right channel and scaled to fit the complete panning range. The energy of the panned signal is kept constant with the source signal. Note that we do not account for directional dependencies of loudness and pitch perception, since we expect the sonification to be used with headphones.

2.5. Frequency Mapping

To map the vertical position component of a segment to frequency, we have to consider different, sometimes opposing objectives. First, we require a linearly perceived increase of frequency for interpretability reasons; while for a pleasing sonification the tone heights of two segments should match consonant intervals. These criteria have to be met under the constraint of a limited frequency spectrum to be used. And finally, frequencies should increase monotonically with a step size of at least the perceptual just notable difference.

To find the most suitable distribution of frequencies Φ (ordered increasing set of fundamental frequencies in Hz) that copes with these competing goals, we formulate a cost function Ψ to be minimized by gradient descent in combination with simulated annealing as follows

$$\Psi(\Phi) = \gamma_l\,\Psi_l(\Phi) + \gamma_d\,\Psi_d(\Phi) + \gamma_o\,\Psi_o(\Phi) + \gamma_r\,\Psi_r(\Phi) \quad (2)$$

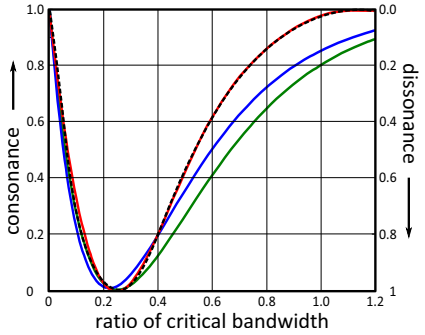


Figure 3: Perceived dissonance of pure tones as a function of the ratio of the critical bandwidth. Experimentally obtained dissonance function by Plomp and Levelt [24] (dashed line), Benson's approximation [25]: $4|x|e^{1-4|x|}$ (green), Sethares' approximation cited in [25] (blue), and our fitting in Equation 5 (red).

with γ_x being a user-defined factor to emphasize particular cost terms Ψ_x that are described in the subsections below. Note that we require the cost terms Ψ_x to be differentiable, since we use gradient descent. Further, we found that an equal distribution of the N fundamental frequencies $\varphi \in \Phi$ in the user-defined frequency range is a suitable initial value to start the gradient descent.

Linear Scaling. The first cost term Ψ_l represents the linearity of the perceived pitches: a property that is important for understanding the conveyed information. To rate the ordered set of fundamental frequencies Φ, we map each of the frequencies $\varphi_i \in \Phi$ (in Hz) to *Zwicker*'s bark scale (critical bandwidth rate, CBR), a perceptual scale of pitches that accounts for the place-spectral analysis of the cochlea [23]:

$$\mathrm{CBR}(\varphi) = 13\,\mathrm{atan}(0.00076\varphi) + 3.5\,\mathrm{atan}(\varphi/7500)^2 \quad (3)$$

As a natural measure of linearity, we take the second (smaller) eigenvalue λ_2 of the 2×2 covariance matrix of the set of vectors

$$\left\{ \begin{pmatrix} \mathrm{CBR}(\varphi_1) \\ 1 \end{pmatrix}, \begin{pmatrix} \mathrm{CBR}(\varphi_2) \\ 2 \end{pmatrix}, \ldots \begin{pmatrix} \mathrm{CBR}(\varphi_N) \\ N \end{pmatrix} \right\}$$

Therefore, we assume that at least a minimum of linearity already exists. Further, we assume the influence of sound pressure level on the perceived pitch to be already compensated by loudness-based amplitude mapping.

Consonant Intervals. To improve acceptance and reduce obtrusiveness and annoyance of our sonification, we account for aesthetics and musicality in terms of consonant intervals. Consonant complex tones exhibit harmonic vibration ratios of their partials (integer multiples) and thus sound pleasant to most people. As measure of consonance of the complex tones of the ordered set of fundamental frequencies Φ (in Hz) with N_H harmonics, we apply the method reported by Plomp and Levelt [24]. The dissonance costs Ψ_d therefore represent the sum over the degree of dissonance of two successive fundamental frequencies $\varphi_i, \varphi_{i+1} \in \Phi$ (in Hz) with their overtones:

$$\Psi_d(\Phi) = \frac{1}{N_H^2(N-1)} \sum_{i=1}^{N-1} \sum_{j,k=1}^{N_H} \mathrm{d}\left(\frac{|j\varphi_i - k\varphi_{i+1}|}{\mathrm{CB}(\sqrt{jk\varphi_i\varphi_{i+1}})} \right) \quad (4)$$

Table 1: Coefficients for sine approximation of dissonance term.

i	α	β	γ
1	2.035	4.340	-1.387
2	3.424	5.662	0.4757
3	1.680	6.469	2.873

The dissonance function d is a perceptual measure that was experimentally derived by Plomp and Levelt [24]. Although several analytical approximations have already been published, we propose a more precise fitting on sine basis (see Table 1 for coefficients, and Figure 3 for a comparison with the original data):

$$\mathrm{d}(x) = \begin{cases} \sum_{i=1}^{3} \alpha_i \sin(\beta_i x + \gamma_i) & , \text{ if } x \leq 1.2 \\ \mathrm{d}(1.2) & , \text{ else} \end{cases} \quad (5)$$

The function $\mathrm{CB}(f_c)$ provides the critical bandwidth of the center frequency $f_c = \sqrt{\varphi\widetilde{\varphi}}$ of the two compared harmonics $\varphi, \widetilde{\varphi}$ according to Zwicker and Terhardt [23]:

$$\mathrm{CB}(f_c) = 25 + 75(1 + 1.4 \cdot 10^{-6} f_c^2)^{0.69} \quad (6)$$

Finally, Ψ_d is normalized to fit the interval $[0, 1]$.

Frequency Order. It is a main requirement of our approach that frequencies in the ordered set Φ increase monotonically. Hence, we have to assure that this criterion is met for all possible solutions of the optimization. The term Ψ_o insures this by penalizing pairs of similar fundamental frequencies in Φ by the sum

$$\Psi_o(\Phi) = \frac{1}{N-1} \sum_{i=1}^{N-1} \left(\frac{0.056\,\mathrm{CB}(\varphi_i)}{\varphi_{i+1} - \varphi_i} \right)^{\alpha} \quad (7)$$

Monotonicity is enforced by the cost function approaching infinity as differences of neighbored frequencies approach zero. Each term of the sum becomes 1 if the frequency differences reach the frequency difference limen, which is about $1/18 \approx 0.056$ times the critical bandwidth [19]. The parameter $\alpha > 0$ is used to adjust the steepness of the function.

Frequency Range. The frequency range available for mapping is limited. Obviously, the human auditory system is restricted to the interval between 20 Hz and about 20 kHz. Furthermore, users may want to narrow this interval even more, for example to the range of musical pitch perception (50 Hz to 5 kHz). The cost term Ψ_r judges the fitness of Φ to match the user-defined frequency interval. Since we presume a monotonic increase in frequency (see section "Frequency Range") , we only have to compare the first and the last fundamental frequency (φ_1, φ_N) with the lower and upper frequency limits (f_l, f_u), respectively. However, we allow the range to exceed these limits at the penalty of rising Ψ_r, represented by sigmoid function terms

$$\Psi_r(\Phi) = \frac{1}{1 + e^{6 + \frac{12(f_u - \varphi_N)}{\mathrm{CB}(f_u)}}} + \frac{1}{1 + e^{6 + \frac{12(\varphi_1 - f_l)}{\mathrm{CB}(f_l)}}} \quad (8)$$

To account for different severities when exceeding the limits at different frequencies (violation of 20 Hz of a limit at 50 Hz is more severe than it is for a limit at 10 kHz), the sigmoidal cost function is scaled to the critical bandwidth (cf. Equation 6) at the particular limit frequency.

3. EVALUATION

We conducted two separate user studies to cover two different purposes. The first user study was conducted during an early stage of development and had a formative character. The goal of such formative evaluation is to provide "insight into which problems occur and why they occur", as well as to provide design feedback [26]. The second user study was designed as a validating user study and conducted in order to evaluate the effectiveness of our sonification approach. The study procedure, as well as the experimental setup, and given tasks were identical for both user studies. However, the participants and the presented auditory stimuli differed between the two user studies. Due to space constraints, we only provide a brief conclusion of the formative user study results here, and include, in exchange, a more detailed discussion on the results of the validating user study.

Experimental Setup. The experiments were conducted in a laboratory insulated from auditive distractions. The audio samples were presented with stereo headphones with volume control.

Stimuli and Tasks. The user study consisted of six sets (**S1** – **S6**) of stimuli and tasks with the purpose to answer different research questions. Auditory stimuli created from video data were presented, without showing the according videos. For **S1** – **S4**, artificial videos with moving textured hexagons were rendered (cf. Figure 4(a)). For **S5** – **S6**, surveillance footage was used (cf. Figure 4(b) and (c)). Stimuli with video data are available at our homepage[2].

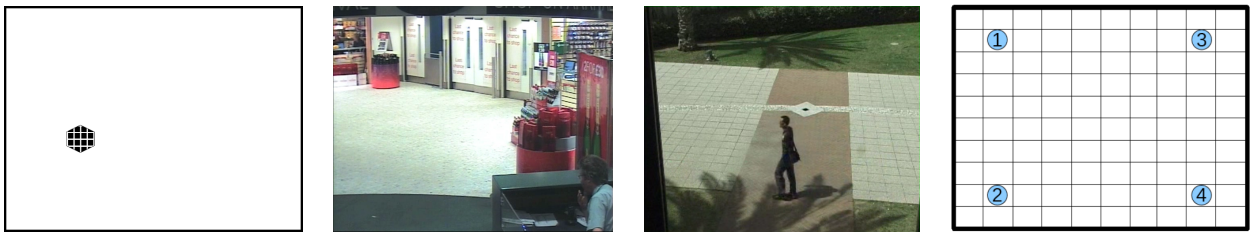
S1: ***Research Question:*** How well can object movement be detected and localized from sonification? (*Accuracy*)
Stimuli: Five stimuli, each with a single moving object. The object movement describes a rhombus, circle, two semicircles with an interruption, an eight, and a triangle.
Task: Sketching trajectories.

S2: ***Research Question:*** How well can similar object movements be distinguished? (*Discrimination*)
Stimuli: Six pairs of stimuli. Each pair consists of two objects with similar movement trajectories presented in succession. The pairs of object movements describe the following patterns: line (back and forth) – with varying slope; circle – var: radius; line (one direction) – var: acceleration; circle – var: object size; rotating object – var: object positions (long distance); rotating object – var: object positions (short distance).
Task: Sketch trajectories.

S3: ***Research Question:*** How sensitive is the sonification to distractors and noise? (*Distraction*)
Stimuli: Three stimuli, each including the movement of a single object. The applied distractors are Gaussian noise (50% normally distributed luminance changes), an image of cluttered background, and MPEG4 coding artifacts (also with cluttered background image).
Task: Sketch trajectories.

S4: ***Research Question:*** Is it possible to detect and distinguish several simultaneously occurring objects? (*Distraction*)
Stimuli: Three stimuli, showing (1) two coexistent objects,

[2] http://www.vis.uni-stuttgart.de/projekte/visual-analytics-of-video-data/sonification.html



(a) (b) (c) (d)

Figure 4: (a) Artificial video showing a hexagonal object); (b) / (c) screenshots of stimuli **S5 / S6** that were provided as context in the user study; (d) template used in the study to sketch recognized trajectories. The blue circles denote the position and order of the calibration objects in the context cue. The grid shows the granularity of the auditory display used in evaluation.

(2) three coexistent objects, and (3) two coexistent objects, where the second appears delayed.
Task: Sketch trajectories.

S5: ***Research Question:*** How well can object movement be detected and localized in real surveillance footage?
Stimuli: One stimulus based on a video from the i-LIDS multi-camera tracking scenario (duration 2:12 min). A contextual image of the video was presented along with the auditory stimuli to facilitate interpretation (cf. Figure 4(b)).
Task: Sketch trajectories.

S6: ***Research Question:*** Does the sonification allow users to detect new and abnormal patterns?
Stimuli: One stimulus based on video [27] showing a pedestrian walk (duration: 8:02 min). Additional to the sonification, a context image was provided to facilitate interpretation (c.f. Figure 4(c)). The first 1:30 min of the stimulus was provided without task in order to learn auditory patterns of normal behavior.
Task: Identification of abnormal behavior.

Study Procedure. First, subjects were asked for basic information, such as their age and profession, followed by an audiometry[3] that took about 5 min. Thereafter, they completed a PowerPoint tutorial (duration $\sim$10 min) that explained the approach and introduced the parameter mappings with the aid of artificial sample videos and their sonifications. After the tutorial, the participants were asked to answer a control question to check whether they understood the technique or not.

Then, we continued with the main evaluation that consisted of the six sets of tasks (**S1** – **S6**) and took about 40 min. Preceding to each stimulus, a *context cue* [16] was provided to enable the participants performing the interpretation tasks. The context cue was the sonification of a calibration pattern that successively showed a rotating textured object at the top left, bottom left, top right, and bottom right. After the context cue, an earcon was played that marked the beginning of the actual stimulus. For **S1** – **S4**, the participants sketched the recognized trajectories on a paper template (cf. Figure 4(d)) while the sonification was played. Acceleration/deceleration had to be marked in green, changes of the object size in red. Further, the trajectories had to be numbered according to their order of appearance. Right after each stimulus, participants had the option to correct and enhance their sketch by drawing the recognized trajectories into a second template.

For **S5**, each recognized trajectory had to be drawn on a separate template, the study operator noted the times when trajectories were identified.

For **S6**, the subjects had to verbally express recognized events. The study operator noted the events including their times.

3.1. Formative User Study

Subjects. Fifteen participants (average age 29.1 years, minimum 27 years, maximum 37 years). Sex was not considered as confounding factor for this study. Twelve participants were students or employees of our university, three participants were professional security guards. Subjects were volunteers and not paid for participation. The audiometry showed that all participants had normal hearing.

Study Results. The formative user study showed that the early version of the sonification was capable of communicating the coarse locations of the objects as well as their trajectories. The study also unveiled that aesthetics and the psychoacoustic of the sonification are critical and have to be taken into account.

3.2. Validating User Study

Subjects. Fourteen participants (average age 32.9 years, minimum 27 years, maximum 57 years). Sex was not considered as confounding factor for this study. Thirteen participants were students or employees of our university. One subject was a physician. Subjects were volunteers and not paid for participation. The audiometry showed that all participants had normal hearing.

Study Results. To judge and compare the accuracy of the sketched trajectories, we consider their start position, end position, and length. The positions are quantized on a lattice with 10 cells for each dimension (x and y, c.f. Figure 4(d)). We chose this granularity according to the expected accuracy and to limit the evaluation effort. We use the Euclidean distance between the cells of the sketched trajectory and the trajectory from ground truth (GT). The distance is normalized to [0,1] by division of the maximum cell distance (i.e., $\sqrt{10^2 + 10^2} \approx 14.14$). To compare lengths between a sketched trajectory and a GT trajectory, we count the transitions between the cells, calculate their difference, and normalize this difference by division with the GT length. A missed trajectory

[3] Applied audiometry: *HTTS-Hörtestprogramm 2.10*. URL: `http://www.sax-gmbh.de/htts/httsmain.htm`

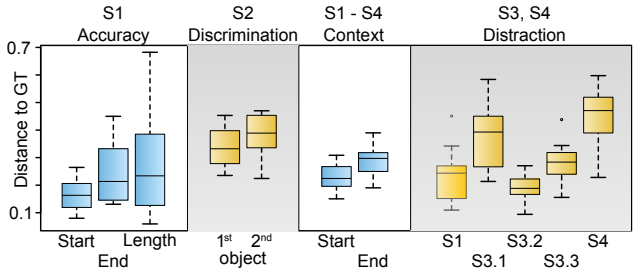


Figure 5: Boxplots of the user study results. Accuracies of the tasks are visualized as relative distances [0,1] to the ground truth. Blue boxplots represent distances of a single parameter (start position, end position, or trajectory length), while yellow boxplots show the combination of the parameters' distances. **S3.**x denotes the x^{th} stimulus of **S3**. First column: general accuracies of particular parameters; second column: accuracies at distinguishing similar object movements; third column: accuracies of start and end positions for all artificial stimuli; fourth column: sensitiveness of the accuracies with respect to distractors.

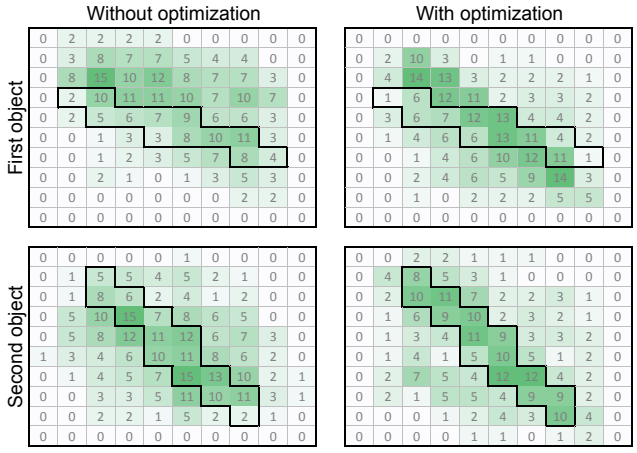


Figure 6: Example of heatmaps for the first pair of stimuli of **S2**. The frequency of how many sketched trajectories traverse a region is mapped to saturation and denoted by the numbers. The black borders denote ground truth trajectories. Left: sonification without optimization, measured during the formative study; right: proposed sonification with optimization; top: first stimulus; bottom: second stimulus with a slightly varied slope.

is penalized with the maximum distance 1 for each parameter. To summarize the accuracies, a combination of the relative distances of the parameters is calculated $\left(\frac{d_{start}+d_{end}+d_{length}}{3}\right)$.

The study results of **S1** – **S4** are depicted in Figure 5. The results of the task and stimuli set **S1** show that localization of the start (median distance: 0.16) and end position (median distance: 0.21) is possible. Moreover, the length of the trajectories can also be estimated roughly (median distance: 0.23).

The results of **S2** show that it is difficult to distinguish similar trajectories. Figure 5 shows that the combined detection accuracies of both the first (median: 0.33) and the second (median 0.39) object of the pair are worse than those of **S1** (median: 0.24). This may have two reasons: First, only a rough localization of a sonified trajectory is possible. Subjects that hear two similar trajectories focus on the movement differences and overestimate them. Second, the context cue is likely to be remembered less accurately for the second object. This is indicated by the worse results of the object appearing second. Another observation made during the study point into the same direction: the accuracy measurements of the start and end positions for all artificial video stimuli **S1** – **S4** (cf. Figure 5 (third column)) exhibit that end positions are generally detected less precisely (median: 0.30) than start positions (median: 0.22).

The localization of trajectories distracted by a background image (**S3.2**, median: 0.19) or a background image with standard MPEG4 artifacts (**S3.3**, median: 0.28) are quite robust (cf. Figure 5, median of **S1** (without distraction): 0.24). Contrary, strong noise (**S3.1**) hinders motion detection and consequently highly interferes with the sonification approach (median: 0.39). Detection of several trajectories simultaneously emerged to be most challenging: sonifying multiple trajectories at the same time drastically reduces localization accuracy (median: 0.47). While most of the subjects detected the existence of two trajectories in **S4.1** and **S4.3** (89%), it was nearly impossible to identify that there were three trajectories present in **S4.2**: only one of the fourteen participants was able to detect it. In **S4**, **S4.3** performed best (median 0.28): it is easier to localize two trajectories when they appear temporally shifted.

Figure 6 shows an example of a heatmap with the results of **S2.1** of the formative study (left) and the validating user study (right) Obviously, sonification of trajectories is more difficult to interpret, if psychoacoustic and aesthetic are not considered. As Figure 6 exhibits, perceptually correct scaling is essential to comprehend the conveyed information. Without the proposed optimization, perception among subjects seems to be more diffuse.

The results of **S5** show that it is – with some limitations – possible to detect and localize trajectories in surveillance footage. The participants were able to sketch most of the trajectories (mean: 0.79, stddev: 0.06) qualitatively correct. It is further possible to detect abnormal behavior (mean: 0.75, stddev: 0.07) due to irregularities in the auditory pattern (**S6**). Moreover, the false detection rate is quite small: on average, there was one false positive detection for each positive example in GT.

Please note that the time the subjects had to learn the standard pattern (1:30 min) as well as the time to learn the video sonification was very short. The effectiveness of the sonification can be expected to be much better when training time increases: it is likely that surveillance operators listening to the sonification for months will be able to identify smaller variations and classify them accordingly.

4. DISCUSSION AND CONCLUSION

In this paper, we introduced a sonification for video data that relies on parameter mapping of quantized optical flow fields. The sonification indicates activity in the video by an abstract sonic pattern with the aim to support situational awareness in the surveillance context. Besides this, we sketched a way to find an optimal balance between the partially opposing goals of an interpretable and aesthetically pleasing sonification. A user study showed that participants are capable of identifying abnormal events by recognizing relevant deviations of the presented soundscape. These results are a requisite to support surveillance operators and indicate that

the proposed sonification can be used as component to support situational awareness. The evaluation also exhibited the limitations of our approach, such as constraints on detection of multiple trajectories or accuracy limits for the estimation of fine movement. A consequence of these results may be the application of such sonification as supportive display.

Future work will extend the mapping by yet neglected psychoacoustic aspects, such as a more sophisticated loudness model that accounts for masking of complex tones. Besides this, optimization of other psychoacoustic aspects should be investigated, such as auditory channel separation, scalability to many displays, and change deafness.

5. ACKNOWLEDGMENTS

This work was funded by German Research Foundation (DFG) by the Priority Program "Scalable Visual Analytics" (SPP 1335).

6. REFERENCES

[1] M. Gill, A. Spriggs, J. Allen, M. Hemming, P. Jessiman, D. Kara, J. Kilworth, R. Little, and D. Swain. (2005) 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., 1999.

[4] 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.

[5] A. Mack, "Inattentional blindness: Looking without seeing," *Current Directions in Psychological Science*, vol. 12, no. 5, pp. 180–184, 2003. [Online]. Available: http://www.jstor.org/stable/20182872

[6] B. Höferlin, M. Höferlin, M. Raschke, G. Heidemann, and D. Weiskopf, "Interactive auditory display to support situational awareness in video surveillance," in *In Proceedings of the International Conference on Auditory Display*, 2011.

[7] C. Nehme and M. Cummings, "Audio decision support for supervisory control of unmanned vehicles," MIT Humans and Automation Laboratory, Cambridge, MA, Tech. Rep. HAL2006-06, 2006.

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

[9] 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.

[10] 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.

[11] A. Johnson and R. Proctor, *Attention: Theory and Practice*. Thousand Oaks, CA: Sage Publications, Inc, 2004.

[12] 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.

[13] P. Meijer, "An experimental system for auditory image representations," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 2, pp. 112–121, 1992.

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

[15] A. Dick and M. Brooks, "Issues in automated visual surveillance," in *Proceeding of VIIth Digital Image Computing: Technique and Applications*, 2003, pp. 195–204.

[16] B. N. Walker and M. A. Nees, "Theory of sonification," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Logos Publishing House, Berlin, 2011, pp. 9–39.

[17] B. Horn and B. Schunck, "Determining optical flow," *Computer Vision*, vol. 17, pp. 185–203, 1981.

[18] F. Grond and J. Berger, "Parameter mapping sonification," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Logos Publishing House, Berlin, 2011, pp. 363–397.

[19] B. C. J. Moore, "Psychoacoustics," in *Springer Handbook of Acoustics*, T. D. Rossing, Ed. Springer Science+Business Media, LLC, New York, 2007, pp. 459–501.

[20] R. Bladon and B. Lindblom, "Modeling the judgment of vowel quality differences," *Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1414–1422, 1981.

[21] Y. Suzuki and H. Takeshima, "Equal-loudness-level contours for pure tones," *The Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 918–933, 2004.

[22] E. Skovenborg and S. Nielsen, "Evaluation of different loudness models with music and speech material," Audio Engineering Society, Tech. Rep., 2004.

[23] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *The Journal of the Acoustical Society of America*, vol. 68, no. 5, pp. 1523–1525, 1980.

[24] R. Plomp and W. Levelt, "Tonal consonance and critical bandwidth," *Journal of the Acoustical Society of America*, vol. 38, no. 4, pp. 548–560, 1965.

[25] D. Benson, *Music: A Mathematical Offering*. New York, USA: Cambridge University Press, 2006.

[26] K. Andrews, "Evaluation comes in many guises," in *AVI Workshop on BEyond time and errors (BELIV) Position Paper*, 2008.

[27] N. Kiryati, T. Raviv, Y. Ivanchenko, and S. Rochel, "Real-time abnormal motion detection in surveillance video," in *19th International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1 – 4.