Reproducing Reality: Multimodal Contributions in Natural Scene Discrimination

OLLI RUMMUKAINEN and CATARINA MENDONÇA, Aalto University

Most research on multisensory processing focuses on impoverished stimuli and simple tasks. In consequence, very little is known about the sensory contributions in the perception of real environments. Here, we presented 23 participants with paired comparison tasks, where natural scenes were discriminated in three perceptually meaningful attributes: movement, openness, and noisiness. The goal was to assess the auditory and visual modality contributions in scene discrimination with short (≤500ms) natural scene exposures. The scenes were reproduced in an immersive audiovisual environment with 3D sound and surrounding visuals. Movement and openness were found to be mainly visual attributes with some input from auditory information. In some scenes, the auditory system was able to derive information about movement and openness that was comparable with audiovisual condition already after 500ms stimulation. Noisiness was mainly auditory, but visual information was found to have a facilitatory role in a few scenes. The sensory weights were highly imbalanced in favor of the stronger modality, but the weaker modality was able to affect the bimodal estimate in some scenes.

CCS Concepts: • **Human-centered computing → Virtual reality**; • **Applied computing → Psychology;**

Additional Key Words and Phrases: Audiovisual, immersive environment, spatial sound, sensory integration, natural scenes

ACM Reference Format:

Olli Rummukainen and Catarina Mendonça. 2016. Reproducing reality: Multimodal contributions in natural scene discrimination. ACM Trans. Appl. Percept. 14, 1, Article 1 (July 2016), 19 pages.
DOI: http://dx.doi.org/10.1145/2915917

1. INTRODUCTION

Human perception is multisensory. To structure the environment from incoming sensory streams robustly and with maximal efficiency, the human brain must merge the information streams to maximize information intake and to increase the reliability of sensory estimates [Ernst and Bülthoff 2004]. The study at hand focuses on early audiovisual processing in natural audiovisual scenes, namely, in the discrimination of scenes in complex environmental attributes from short (≤500ms) scene exposures. So far, most studies on multisensory processing have used impoverished stimuli (lights and noise bursts) and simple detection or localization tasks. Here, for the first time, a controlled study is implemented to

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 240453 and from the Academy of Finland project Audiovisual Space no. 266239. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant no. 659114.

Authors' addresses: O. Rummukainen and C. Mendonça, Aalto University School of Electrical Engineering, Otakaari 5 A, 02150 Espoo, Finland; emails: {olli.rummukainen, catarina.mendonca}@aalto.fi.

DOI: http://dx.doi.org/10.1145/2915917

test natural-scene discrimination with complex perceptual attributes and with immersive audiovisual stimulation.

In a real-world environment, most events generate stimulation to multiple senses. For example, when thinking about localizing a moving object, we can both see a car moving across a street and hear the sound emanating from the car's engine move in space. Through the integration of such partially redundant data, the human brain is capable of compensating for modality-specific disturbances, for example, moment-to-moment variability in the sensory stimulation due to, in the case of localizing a car, competing engine sounds from other nearby vehicles [De Gelder and Bertelson 2003]. Spatial and temporal coincidence of sensory inputs are typically needed for sensory integration, but also the semantic content of the stimulus affects the integration process, especially when using meaningful stimuli [Spence and Squire 2003; Laurienti et al. 2004; Doehrmann and Naumer 2008].

Semantically congruent auditory cues have been found to aid in identifying occluded visual objects [Chen and Spence 2010] and to modulate visual awareness in a binocular rivalry task [Chen et al. 2011]. Recently, evidence for the importance of semantic congruence in audiovisual scene processing was provided by Tan and Yeh [2015], who showed that congruent sound scene facilitates unconscious visual scene processing when the visual scene is masked. Their findings expand the importance of auditory cues from identifying individual objects to full scenes. A similar unconscious facilitation effect has been found with semantically meaningless signals, in which cross-modal influences have been found to facilitate the detection of objects or events [Evans and Treisman 2010; Pérez-Bellido et al. 2013].

Additionally, audiovisual sensory integration has been shown to be beneficial for human observers in multiple tasks. We are capable of synchronizing our actions with the environment more robustly when given audiovisual stimuli instead of unimodal stimuli [Armstrong and Issartel 2014]. Sensory discrimination is finer with bimodal signals than with unimodal signals [Koene et al. 2007], and the effect is more robust when given meaningful stimuli, such as biological motion [Mendonça et al. 2011]. Visual processing of looming signals is enhanced when accompanied by a related looming auditory signal, and especially when using naturalistic stimulus structure [Conrad et al. 2013]. Similarly, a nominally visual event of perceiving a ball's path is aided by naturalistic auditory cues [Ecker and Heller 2005]. However, cross-modal influences can also bias the unimodal estimates of environmental properties such as the speed of motion [López-Moliner and Soto-Faraco 2007].

A significant amount of research has been devoted to understanding the mechanisms of sensory integration. In many cases, the brain seems to function in a near-optimal manner, in which the unimodal sensory estimates are weighted proportionally to their inverse variances and integrated into a multimodal estimate [Treisman 1998; Cheng et al. 2007; Wozny et al. 2008]. Another theory, sensory dominance, was discovered in detection tasks, in which visual modality was shown to dominate the detection of light-tone blinks [Colavita 1974] and has ever since been identified in a wide range of multisensory tasks. However, as identified in recent research [Alais and Burr 2004; Hecht and Reiner 2009; Yuval-Greenberg and Deouell 2009], sensory dominance is not a rule for dealing with multisensory stimuli; rather, it is a result of multisensory integration when there are very imbalanced sensory weights. A third theory, sensory averaging, forms the multisensory estimate by simply averaging the unimodal estimates without assuming anything about the unimodal reliabilities [Treisman 1998].

The goal of the current study is to investigate sensory integration in a naturalistic setting and with more complex tasks than in earlier studies. As the dependent variable, we observe the probability of successful discrimination of reproduced real-world environments in *movement*, *openness*, and *noisiness*. As independent variables, we have three settings affecting the sensory information given by the system: sensory modality, stimulus duration, and stimulus scene. We vary the reproduction condition between unimodal, auditory or visual, or bimodal audiovisual stimulation. We test the time required for information processing by using three stimulus durations, 100ms, 200ms, and 500ms. Finally, we have 8 perceptually different scenes for which the discrimination is evaluated.

We hypothesize that the probability of successful scene discrimination is enhanced in the bimodal condition when compared to either of the unimodal conditions, and that the success of discrimination varies between scenes. In addition, we hypothesize both modalities to have an effect on all three attributes instead of the stronger modality always capturing the bimodal estimate. Last, we assume the unimodal conditions to require longer stimulation to yield discrimination success comparable to the bimodal condition.

The results will provide guidelines for content creation for immersive systems, and for the development and evaluation of multimodal technology by shedding light on the early processing of congruent auditory and visual information in real-world environments, and their relative importance when evaluating meaningful natural-scene attributes. The stimulus scenes are published online as spherical video files and A-format spatial audio recordings along with the data from the perceptual experiment at http://research.spa.aalto.fi/publications/papers/natural-scene-discrimination/.

2. MATERIALS AND METHODS

2.1 Ethics Statement

The stimulus material and experimental setup were approved by the ethics review board of Aalto University. All the participants gave written informed consent prior to participating in the experiment.

2.2 Participants

A total of 23 participants took the test. The participants were naïve with regard to the goal of the study. Nine of the participants were female, and the average age was 26.5y ($SD = 6.6$). All participants reported to have normal or corrected-to-normal vision and normal hearing. No acuity screening or audiogram measurements were considered necessary, as the participants were supposed to perceive and assess the reproduced environments as they would experience natural situations in their everyday lives. The participants received extra course points as a compensation for their contribution. The authors did not participate in the experiment. One of the participants was excluded from the final analysis due to a distinctively large number of missing answers.

2.3 Catalog of Environments

The stimulus scenes were chosen based on our previous study focusing on perceptual similarity of natural scenes [Rummukainen et al. 2014]. The eight scenes chosen here were found to span a two-dimensional perceptual space, with meaningful perceptual gradients related to eventfulness of the scene (i.e., the amount of temporal variation in both the sound scene and visual scene) and the perceived level of openness or expansion of the scene. In this study, we divide the eventfulness into two components: movement and noisiness. Screen captures of the stimulus scenes are presented in Figure 1. The scenes are recorded in the Helsinki capital region in Finland during the summertime. The scenes can be previewed online at http://research.spa.aalto.fi/publications/papers/natural-scene-discrimination/. In addition, full-quality spherical videos and A-format audio files are provided.

Classifying natural scenes is quite difficult; thus, we present the scenes in two tables. Table I summarizes the scenes as objectively as possible and Table II lists the perceptual concepts most closely related to these stimuli based on results presented in Rummukainen et al. [2014]. In Table I, the scenes are classified based on whether they are indoor or outdoor scenes (setting), how much the scene contains movement (movement), and is the sound scene noisy or quiet (sound scene). The sound scene classification is accompanied by an A-weighted sound pressure level measurement, but the label (noisy

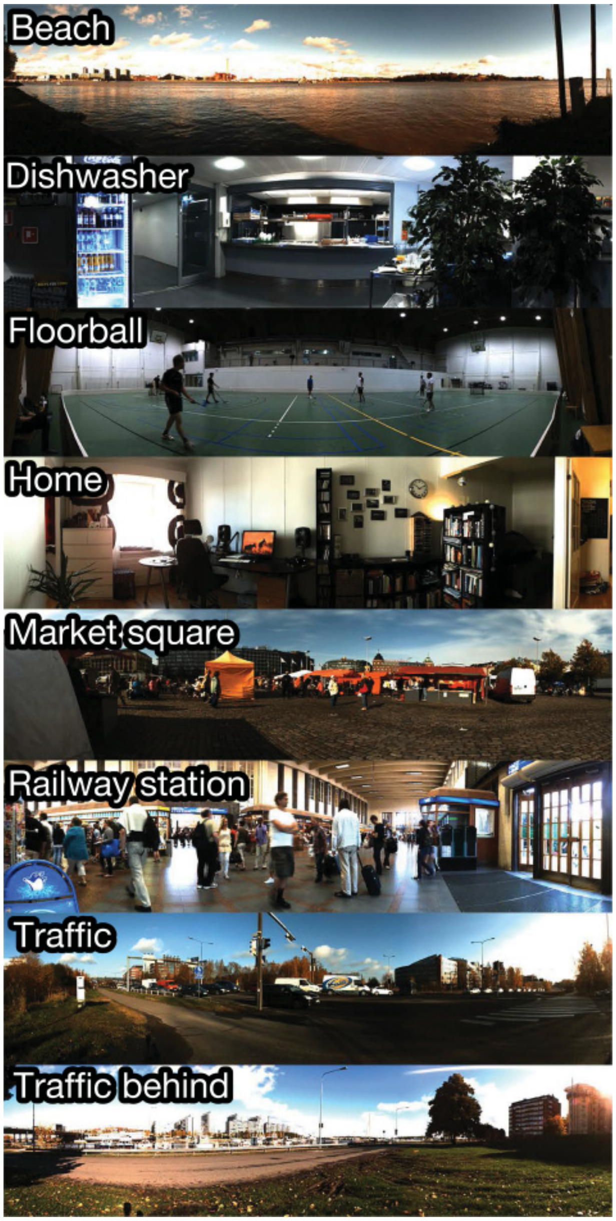

Fig. 1. Collage of the stimulus contents. The scenes can be previewed and downloaded online at http://research.spa.aalto.fi/publications/papers/natural-scene-discrimination/.

vs. quiet) is determined based on perceptual judgment. The reason for this is that the A-weighted SPL measurement does not accurately reflect how noisy the sound scene is, for example, in the case of #*Floorball* vs. #*Traffic behind*, the SPL measurement is the same, but the latter contains traffic noise and the former temporally separated sound events that are not perceptually as noisy as the continuous sound from the traffic.

Table I. Properties of the Scenes and a Description of the Content

Name	Setting	Movement	Sound scene ($L_{Aeq,10s}$)	Description
Beach	Outdoor	None	Quiet, 41dB	Expansive view of the sea and quiet
Dishwasher	Indoor	None	Noisy, 63dB	Close-up view of a dishwasher
Floorball	Indoor	Plenty	Quiet, 59dB	Indoor hall with a game of floorball
Home	Indoor	None	Quiet, 41dB	Small room with TV on
Market square	Outdoor	Plenty	Quiet, 58dB	Market square with a few stalls and people
Railway station	Indoor	Plenty	Noisy, 62dB	Large and busy departure hall
Traffic	Outdoor	Plenty	Noisy, 71dB	Busy street in front of the viewer
Traffic behind	Outdoor	None	Noisy, 59dB	Busy street behind the viewer

Table II. Perceptual Concepts Related to the Chosen Stimulus Scenes in Rummukainen et al. [2014]

Name	Perceptual concepts
Beach	Calm, Quiet, Pleasant
Dishwasher	Background sound, noisy, movement
Floorball	Large indoor space, Enclosed space, People, Echo, A lot to attend
Home	Quiet, Pleasant
Market square	Movement, Conversation, Background sound, Noisy
Railway station	Echo, A lot to attend, People
Traffic	Noisy, Traffic
Traffic behind	Outdoor space, Noisy, Traffic

In Table II, the presented concepts result from semistructured interviews, in which the participants were allowed to use their own words to describe the audiovisual scenes. Therefore, Table II captures the whole range of possible perceptual attributes, and the reader should notice that none of the scenes can be described purely in terms of the three perceptual attributes (movement, openness, and noisiness) applied in this study.

2.4 Reproduction

The experiment was conducted in an immersive audiovisual environment, depicted in Figure 2, located at the Department of Signal Processing and Acoustics in Aalto University School of Electrical Engineering. The environment is built inside of an acoustically treated room and consists of three high-definition video projectors producing a horizontal field-of-view of 226° at the viewing position on three acoustically nearly transparent screens. Vertical field-of-view is 57° at the center of each screen. The screens are 2.5 × 1.88m each and installed to follow the shape of the base of a pentagon. The display area extends to the ground. Distance from the observation position to the center of each screen is 1.72m. More detailed information about the technical specifications can be found in Gómez Bolaños and Pulkki [2012].

The videos were captured with a recording device capable of producing a spherical video (Point Grey Research: Ladybug 3). The videos were cropped to reproduce a 226° slice of the full circle on the screen, making the visual scene consistent with the auditory scene. The video was recorded and reproduced at 16 frames per second and the resolution of the final video was 4320 × 1080 pixels produced by the three projectors. With this resolution and viewing distance, the interpixel distance is 3.5arcmin, which

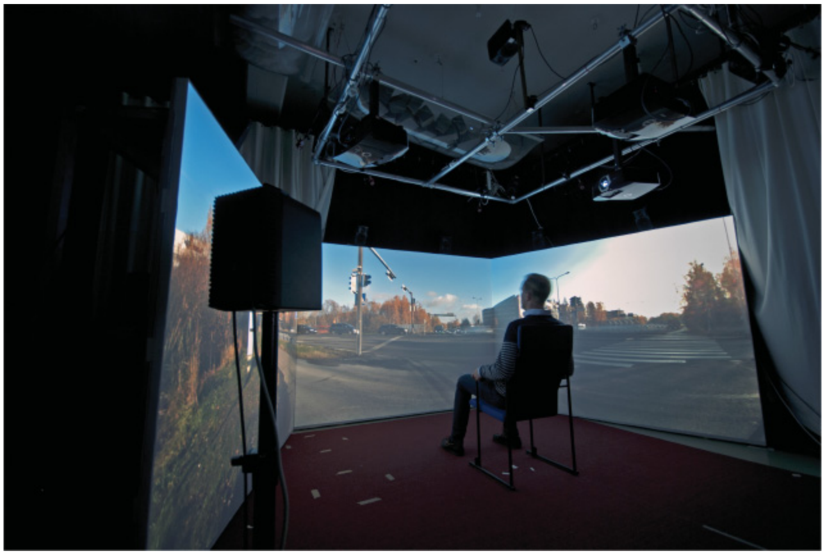

Fig. 2. View of the audiovisual environment with a participant seated in the observation position. The projectors can be seen overhead of the participant. The projection screens cover the frontal loudspeakers.

affects the perceived sharpness of the image, since humans with normal visual acuity can discriminate two lines separated by 1arcmin.

The audio reproduction system consisted of 29 loudspeakers, which were located on a sphere with a 2.1m radius centered at the observation position. The loudspeaker layout and the projection screen setup are depicted in Figure 3. The signals to the loudspeakers were derived with Directional Audio Coding (DirAC; Pulkki [2007]; Politis and Pulkki [2011]), which is a recently proposed parametric spatial audio technique. DirAC analyzes and synthesizes the sound field from A-format microphone (Soundfield SPS200) signals recorded from a real-world environment; the reproduction results in a perceptually good correspondence with the original sound scene [Vilkamo et al. 2009].

2.5 Procedure

2.5.1 *Overview.*

The test was composed of paired-comparison tasks of natural scenes in three perceptual attributes. The task was to choose the first or the second stimulus scene of each pair, that was perceived as having "more" of the perceptual attribute in question, in a 2-alternative forced choice design. There were three tested attributes: *openness*, *noisiness*, and *movement*. The participants were given a chance to practice discriminating the scenes in each of the attributes and reproduction conditions before the actual test began. It was emphasized that the attribute estimation should be based on all available sensory information given by the system, that is, in the case of audiovisual stimulation, both the sound and visuals should be considered. The attribute selection was based on our previous study [Rummukainen et al. 2014], in which these attributes were identified as the most perceptually meaningful attributes to naïve perceivers, when they were asked to categorize natural scenes according to perceived similarity.

2.5.2 *Baseline Scaling.*

To reduce the length of the study, a stimulus scene was always compared to a reference instead of having a full paired-comparison design with the 8 scenes. The reference scenes were determined in a baseline scaling session, and the reference scene was different for each attribute. In the baseline scaling session, stimulus scenes were scaled in a paired-comparison design according

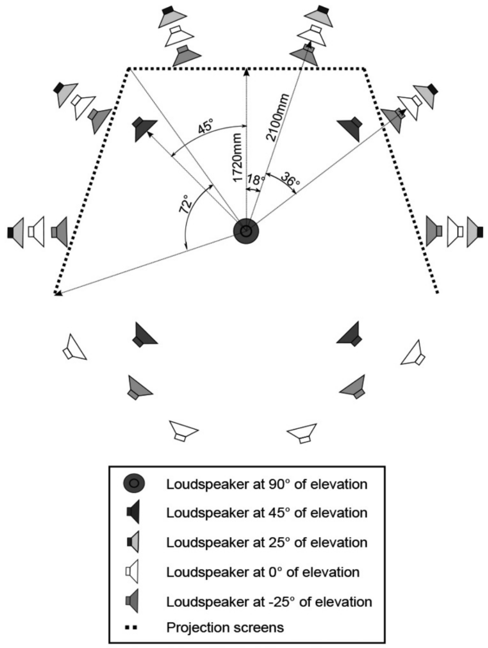
45°
2100mm
1720mm
18°
36°
72°
Loudspeaker at 90° of elevation
Loudspeaker at 45° of elevation
Loudspeaker at 25° of elevation
Loudspeaker at 0° of elevation
Loudspeaker at -25° of elevation
Projection screens

Fig. 3. Loudspeaker setup and projection screens. The shading denotes elevation of the loudspeakers.

to the three perceptual attributes by three participants (1 female, 2 male), who did not take part in the main test. The goal was to find three scenes with intermediate values for each of the three attributes so that comparing the remaining 7 scenes to the reference would be meaningful. If the chosen reference for noisiness would be the noisiest of the 8 scenes, for example, then the comparison task would become trivial and it would be easy to construct an answering strategy. The secondary goal was to obtain an approximate rank ordering of the scenes in each attribute to facilitate interpreting and visualizing our results.

The baseline scaling experiment was a full paired-comparison design with three different stimulus durations (100ms, 200ms, and 500ms) and audiovisual reproduction only. Each pair of stimuli was evaluated once with every duration combination, resulting in nine evaluations for each pair of contents. Counting all possible pairs, there were 28 different combinations of the stimulus contents, adding up to 252 paired comparisons for each of the three perceptual attributes.

2.5.3 *Main Test.* Each participant evaluated two of the three attributes in two separate sessions, each session including only one attribute under test. The two sessions could also be combined into one long session depending on the participant's preference; the attributes would still be evaluated separately, however. The attributes are later analyzed in three separate analyses. Each attribute was evaluated in three different durations, and in three different reproduction conditions—audio only (A), visual only (V), and audiovisual (AV)—in a blocked design, in which the reproduction condition was constant within one block. In one block, there were three different presentation durations, either pairs of 100ms, 200ms, or 500ms. The remaining 7 scenes were compared to the selected reference 12 times for each presentation duration and reproduction condition to add more reliability to the results. Thus, one block consisted of (3 durations) × (7 scene pairs) × (12 repetitions), and one session consisted of 3 blocks with A, V, or AV reproduction. To give an example, one block might evaluate noisiness

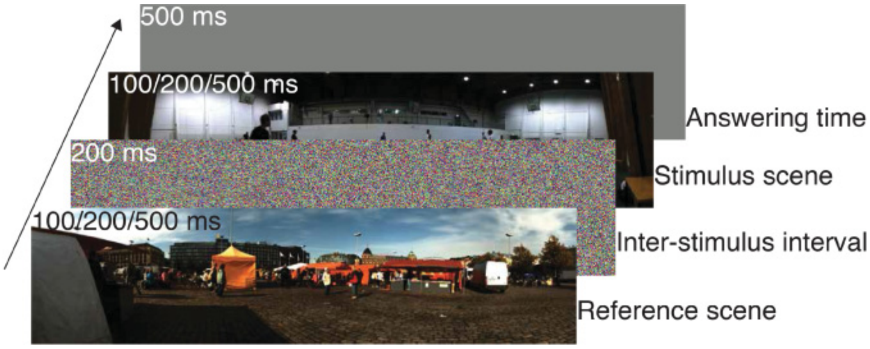

Fig. 4. Structure of one paired comparison task. The reference scene was displayed first, followed by a static pixel noise frame to clear the short-term visual memory. After 200ms, the stimulus scene was presented and followed by an answering time of 500ms. The reference scene and the stimulus scene were always equally long in duration. In the audio-only condition, there was a mid-gray frame presented during the whole block.

with audio-only reproduction in three different duration conditions. There was a brief break after each block. This was repeated for each attribute.

Within a paired comparison, there was a 200ms interstimulus interval and an answering time of 500ms after the pair. If the participant failed to answer during the 500ms period, that pair was marked as missing response. The structure of one paired comparison task is presented in Figure 4. In sum, for each participant, the test was divided into two sessions, both including three blocks according to the reproduction conditions, and each block including three stimulus durations. As a complete design we had (3 attributes) × (3 reproduction conditions) × (3 stimulus durations) × (7 scene pairs) × (12 repetitions).

The reason for short stimulation and short answering time was to prevent the participants from easily memorizing the scenes and forcing them to actively process the scenes in every pair. In addition, the rapid progress in every block prevented the participants, presented with reproduced real-world scenes, from associating the scenes with some personal experiences or memories that they might have had regarding the presented spaces.

In total, we had 252 paired comparisons for each block. The duration of one block was 9min, and each participant evaluated six blocks either in one session or divided into two sessions depending on one's preference. The presentation order of the attributes, blocks, and pairs was randomized. Taking training session and the breaks into account, total duration of the test was approximately 80min.

3. RESULTS

3.1 Baseline Scaling

The baseline data with different stimulation durations were pooled because there were no significant differences in the scalings between the different duration conditions. We computed Fleiss's Kappa for interrater agreement. The resulting kappa values are: Movement 0.60, Openness 0.73, and Noisiness 0.80. The values can be considered acceptable [Landis and Koch 1977]. Figure 5 presents the scaling results in the three attribute dimensions: movement, openness, and noisiness. Respectively, the chosen reference scenes, *#traffic behind*, *#railway station*, and *#market square*, are presented in Figure 5 in relation to other scenes. The reference scenes were chosen based on their approximate scaling to the mid-level in the corresponding attribute dimensions. In Figures 6, 7, and 8, the order of the scenes is defined by the baseline scaling in the respective attributes.

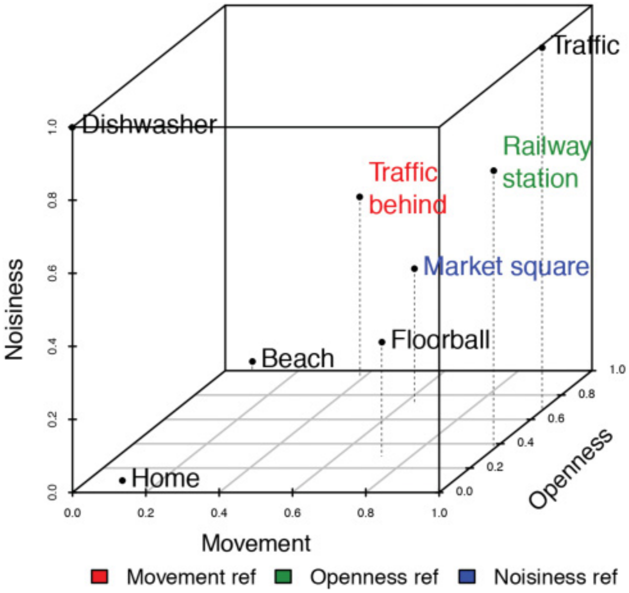

Fig. 5. Baseline scaling of the stimuli in the three attribute dimensions. The chosen reference scene in each attribute dimension is denoted by color. They are as follows: Movement—#*Traffic behind*, Openness—#*Railway station*, and Noisiness—#*Market square*.

3.2 Probabilities of Successful Discrimination

3.2.1 *Analysis Methods.* Each participant evaluated two of the three attributes; here, the attributes are analyzed independently even though the test was not completely a between-subjects design. The decision to have each participant evaluate two attributes instead of having three groups of participants was made to increase the amount of data. The attributes under investigation are perceptually highly different from each other and were evaluated in separate sessions, reducing the dependency between the attributes. Probability of successful attribute discrimination was inspected in the three sensory modalities and three durations.

We performed two different analyses on the data: a generalized linear mixed-model analysis of the discrimination success in all combinations of modality and duration, and a by-scene ANOVA. All analyses were conducted in the R language and environment [R Core Team 2015]. Mixed-model analysis was performed by functions in the R-package *lme4* [Bates et al. 2015]. In the mixed-model analysis, we entered modality and duration (with interaction term) as fixed effects into the model. As random effects, we had intercepts for participants and scenes, as well as by-participant and by-scene random slopes for the effects of modality and duration. P values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question. The mixed model is able to give the odds $(x : 1)$ of successful discrimination in the different conditions, while taking into account the variation caused by the participants and scenes. We classified successful discrimination based on rank ordering of the scenes from the baseline scaling session. Furthermore, the model tells how the baseline success varies in the random effects, that is, whether the scenes are equally easy to discriminate and whether the participants are equally good at the task.

Regarding the ANOVA, each participant compared every scene to the reference 12 times, which produced an individual probability score of perceiving a scene as having "more" of some attribute than

Table III. Significance of Fixed Effects in Movement Discrimination, and Significant Odds Effects

Fixed effects	**Likelihood test statistic**	**Odds effects**
Modality	$X^2(2) = 8.47$, $p = 0.01$	A: 1.93 ± 1.50 (s.e.)
		V: 7.56 ± 1.27 (s.e.)
		AV: 5.17 ± 1.34 (s.e.)
Duration	$X^2(2) = 1.68$, $p = 0.43$	
Modality × *Duration*	$X^2(4) = 4.91$, $p = 0.30$	
Random effect: Scene	**Intercept**	
Beach	16.39	
Floorball	2.12	
Home	26.88	
Dishwasher	3.82	
Market square	1.64	
Railway station	3.96	
Traffic	4.30	

the reference. This analysis is not dependent on the accuracy of the rank ordering of the scenes. In Figures 6, 7, and 8, the scenes positioned below the reference in the baseline scaling should have $P(more) = 0$ probability of being perceived as having more movement, openness, or noisiness. Similarly, scenes above the reference should have more of the attributes with $P(more) = 1$, in the case of perfect discrimination. Within each scene, a two-way ANOVA (Modality × Duration) was conducted, and significant main effects were inspected by Tukey's pairwise post-hoc tests. Mauchly's sphericity test was performed in all cases and, when required, the Greenhouse-Geisser correction was applied to meet the assumption of sphericity.

3.2.2 *Movement.*

A total of 15 participants evaluated the movement attribute. Table III summarizes the results from the mixed-model analysis for the movement attribute. Only modality was found to be a significant fixed effect in the model, with the odds of successful discrimination being the highest for V stimulation, followed by AV stimulation. The random effect caused by the scene shows large variation: the baseline odds of getting the discrimination correct with the #*Home* scene is 26.9:1, while the odds for the #*Market square* scene is 1.6:1. This variation leads to the need to inspect the scenes individually.

In Figure 6, the scenes are ordered with their baseline movement (from the baseline scaling session) increasing from left to right. Each scene was compared to the reference scene, #*Traffic behind*, 12 times by each participant, which produced a mean probability score for each participant. Significant main effects and post-hoc tests are summarized in Table IV. In movement discrimination, 102 pairs were missing an answer, corresponding to 0.9% of the full dataset.

3.2.3 *Openness.*

A total of 14 participants evaluated the openness attribute. Table V summarizes the results from the mixed-model analysis for the openness attribute. Only modality was found to be a significant fixed effect in the model, with the odds of successful discrimination being the highest for V and AV stimulation. The random effect caused by the scene shows the scene #*Dishwasher* to be by far the easiest to discriminate, with odds of 141.5:1.

Looking at the scenes individually, in Figure 7, the scenes are ordered with their baseline openness (from the baseline scaling session) increasing from left to right. Each scene was compared to the reference scene: #*Railway station*. The significant effects are summarized in Table VI. In openness discrimination, 63 pairs were missing an answer, corresponding to 0.6% of the full dataset.

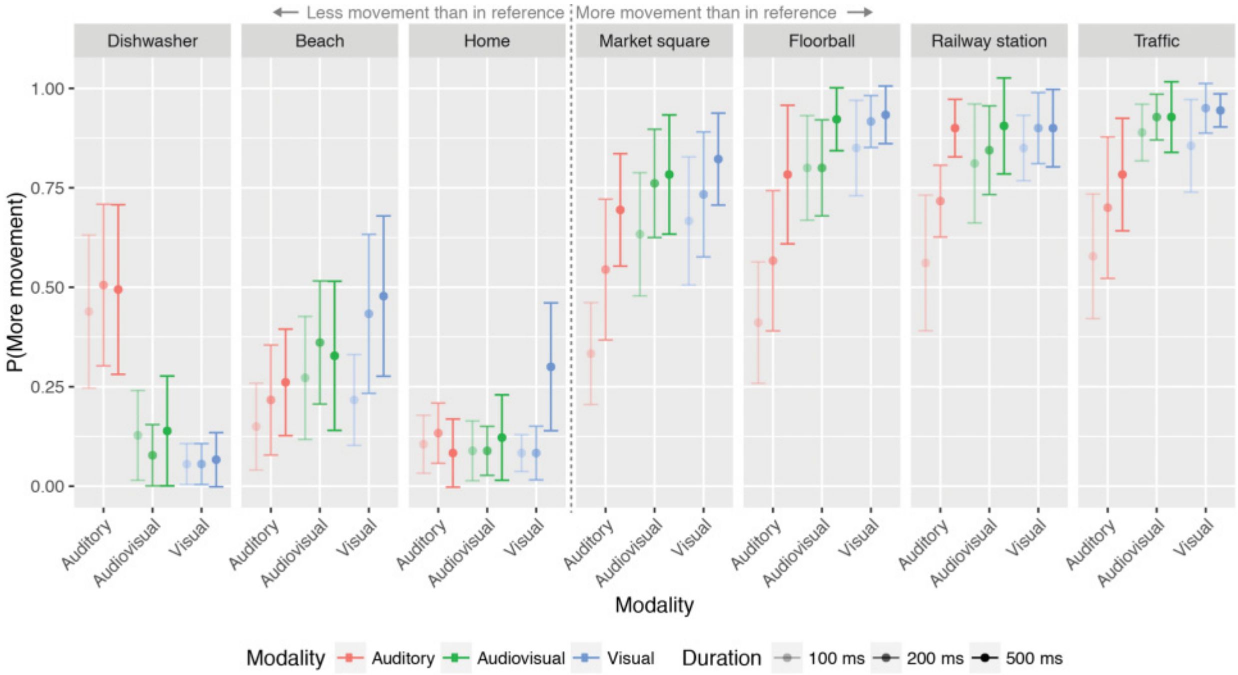

Fig. 6. Movement discrimination task between the stimulus scenes and the reference (*#Traffic behind*). The amount of baseline movement is increasing from left to right, and the reference scene is located between the 3rd and 4th scenes. Each sample mean results from 180 paired comparisons made by 15 participants evaluating the same pair 12 times each. The bars denote the 95% confidence intervals of the mean calculated from the 15 mean movement scores and assuming normal distribution.

Table IV. Significant Main Effects in Movement Discrimination and Post-Hoc Tests with $p < 0.05$

Dishwasher:	*Modality* $F_{(2,28)} = 17.13$, $p < 0.01$	A > V & AV
Beach:	*Modality* $F_{(2,28)} = 3.43$, $p = 0.05$	V > A
	Duration $F_{(2,28)} = 8.02$, $p < 0.01$	500ms > 100ms
Home:	*Modality* × *Duration* $F_{(4,56)} = 3.91$, $p = 0.03$	$V_{500} > V_{100,200}$
	$Duration_V$ $F_{(2,28)} = 7.08$, $p = 0.01$	
Market square:	*Modality* $F_{(2,28)} = 5.67$, $p = 0.01$	AV & V > A
	Duration $F_{(2,28)} = 15.42$, $p < 0.01$	200ms & 500ms > 100ms
		AV_{100} & $V_{100} > A_{100}$
		$A_{500} > A_{100}$
Floorball:	*Modality* × *Duration* $F_{(4,56)} = 5.20$, $p < 0.01$	$AV_{100,200}$ & $V_{100,200} > A_{100,200}$
	$Modality_{100}$ $F_{(2,28)} = 18.64$, $p < 0.01$	$A_{500} > A_{100}$
	$Modality_{200}$ $F_{(2,28)} = 13.84$, $p < 0.01$	
	$Duration_A$ $F_{(2,28)} = 13.4$, $p < 0.01$	
Railway station:	*Modality* × *Duration* $F_{(4,56)} = 8.58$, $p < 0.01$	$V_{100,200} > A_{100,200}$
	$Modality_{100}$ $F_{(2,28)} = 6.51$, $p = 0.02$	$A_{500} > A_{100,200}$
	$Modality_{200}$ $F_{(2,28)} = 6.48$, $p = 0.02$	
	$Duration_A$ $F_{(2,28)} = 15.68$, $p < 0.01$	
Traffic:	*Modality* $F_{(2,28)} = 8.93$, $p < 0.01$	AV & V > A
	Duration $F_{(2,28)} = 11.55$, $p < 0.01$	500ms > 100ms

Table V. Significance of Fixed Effects in Openness Discrimination, and Significant Odds Effects

Factor	Likelihood test statistic	Odds effects
Modality	$X^2(2) = 12.11, p = 0.002$	A: 0.87 ± 1.66 (s.e.) V: 5.53 ± 1.37 (s.e.) AV: 5.03 ± 1.93 (s.e.)
Duration	$X^2(2) = 2.89, p = 0.23$	
Modality × *Duration*	$X^2(4) = 1.74, p = 0.78$	
Random effect: Scene	**Intercept**	
Beach	9.56	
Floorball	5.04	
Home	12.97	
Dishwasher	141.50	
Market square	12.96	
Traffic	1.37	
Traffic behind	5.44	

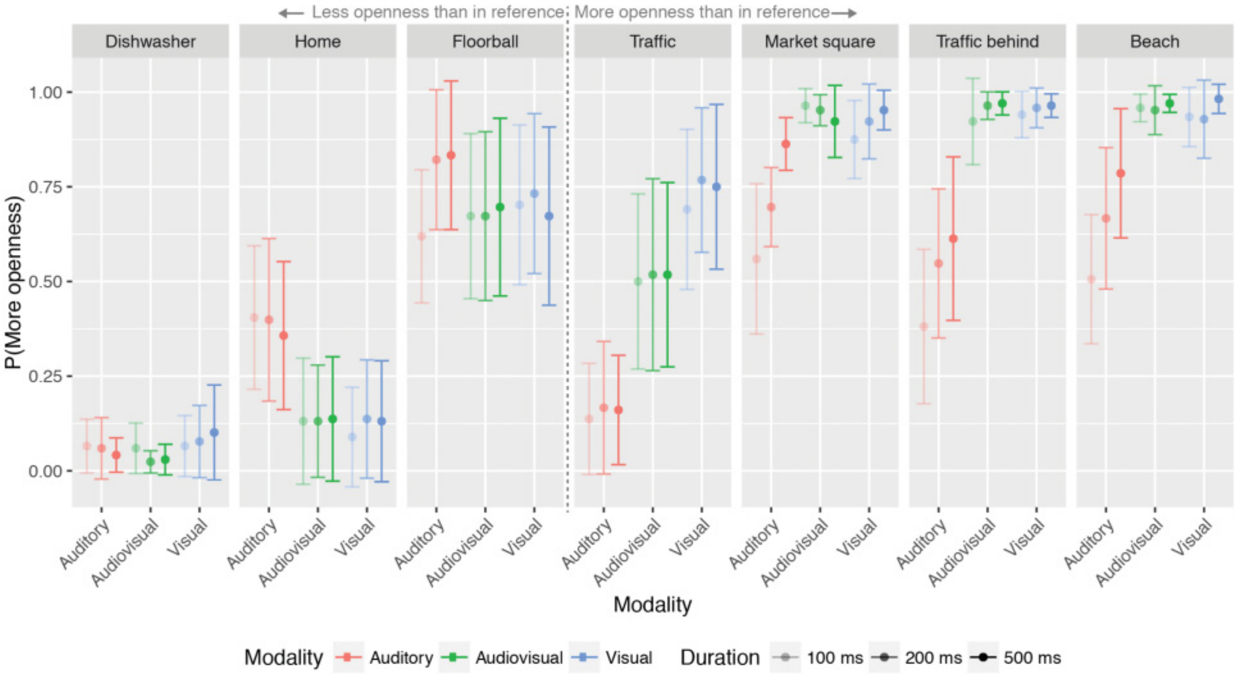

Fig. 7. Openness discrimination task between the stimulus scenes and the reference (#*Railway station*). The amount of baseline openness is increasing from left to right, and the reference scene is located between the 3rd and 4th scenes. Each sample mean results from 168 paired comparisons made by 14 participants evaluating the same pair 12 times each. The bars denote the 95% confidence intervals of the mean calculated from the 14 mean openness scores and assuming normal distribution.

3.2.4 *Noisiness.* A total of 15 participants evaluated the noisiness attribute. The mixed-model analysis for noisiness discrimination does not yield any significant fixed effects, but again the by-scene random variation is substantial in Table VII. Therefore, the scenes are inspected individually. In Figure 8, the scenes are ordered with their baseline noisiness increasing from left to right. Each scene

Table VI. Significant Main Effects in Openness Discrimination and Post-Hoc Tests with $p < 0.05$

Dishwasher:	-	-
Home:	*Modality* $F_{(2,26)} = 10.76$, $p < 0.01$	A > AV & V
Floorball:	*Modality* × *Duration* $F_{(4,52)} = 8.79$, $p < 0.01$	$A_{200,500} > A_{100}$
	$Duration_A$ $F_{(2,26)} = 11.37$, $p < 0.01$	
Traffic:	*Modality* $F_{(2,26)} = 20.41$, $p < 0.01$	V > AV > A
Market square:	*Modality* × *Duration* $F_{(4,52)} = 7.01$, $p < 0.01$	$A_{500} > A_{100,200}$
	$Duration_A$ $F_{(2,26)} = 9.41$, $p < 0.01$	$V_{500} > V_{100}$
	$Duration_V$ $F_{(2,26)} = 3.77$, $p = 0.05$	$AV_{100,200}$ & $V_{100,200} > A_{100,200}$
	$Modality_{100}$ $F_{(2,26)} = 12.86$, $p < 0.01$	
	$Modality_{200}$ $F_{(2,26)} = 14.20$, $p < 0.01$	
Traffic behind:	*Modality* $F_{(2,26)} = 24.24$, $p < 0.01$	AV & V > A
	Duration $F_{(2,26)} = 7.05$, $p < 0.01$	500ms > 100ms
Beach:	*Modality* × *Duration* $F_{(4,52)} = 9.10$, $p < 0.01$	$A_{500} > A_{100}$
	$Duration_A$ $F_{(2,26)} = 9.95$, $p < 0.01$	AV & V > A
	$Modality_{100}$ $F_{(2,26)} = 29.46$, $p < 0.01$	
	$Modality_{200}$ $F_{(2,26)} = 12.59$, $p < 0.01$	
	$Modality_{500}$ $F_{(2,26)} = 5.94$, $p = 0.03$	

Table VII. Significance of Fixed Effects in Noisiness Discrimination, and Significant Odds Effects

Factor	Likelihood test statistic	Odds effects
Modality	$X^2(2) = 0.27$, $p = 0.87$	
Duration	$X^2(2) = 0.11$, $p = 0.95$	
Modality × *Duration*	$X^2(4) = 7.20$, $p = 0.12$	
Random effect: Scene	**Intercept**	
Beach	15.52	
Floorball	2.47	
Home	11.69	
Dishwasher	11.87	
Railway station	6.17	
Traffic	39.56	
Traffic behind	3.64	

was compared to the reference scene #*Market square* 12 times. The significant effects are summarized in Table VIII. In noisiness discrimination, 76 pairs were missing an answer, corresponding to 0.7% of the full dataset.

3.3 Adaptation Effects

Each participant evaluated each trial 12 times, which may have resulted in adaptation effects and improved discrimination accuracy in late repetitions. We inspected how the answering patterns evolved in each scene through the repetitions in all conditions and attributes. This was done by dividing the data according to the number of repetitions so that the first six repetitions in each condition combination for each participant formed one set of data and the last six formed another set of data. It is important to note that, within the attribute sessions, the A, V, and AV modalities were presented in random order, and, similarly, the durations were scrambled. Therefore, the first six repetitions of a given modality/duration combination occur at different points in time for different participants. Table IX summarizes a two-way repeated measures ANOVA (Repetition × Scene) in which the data is

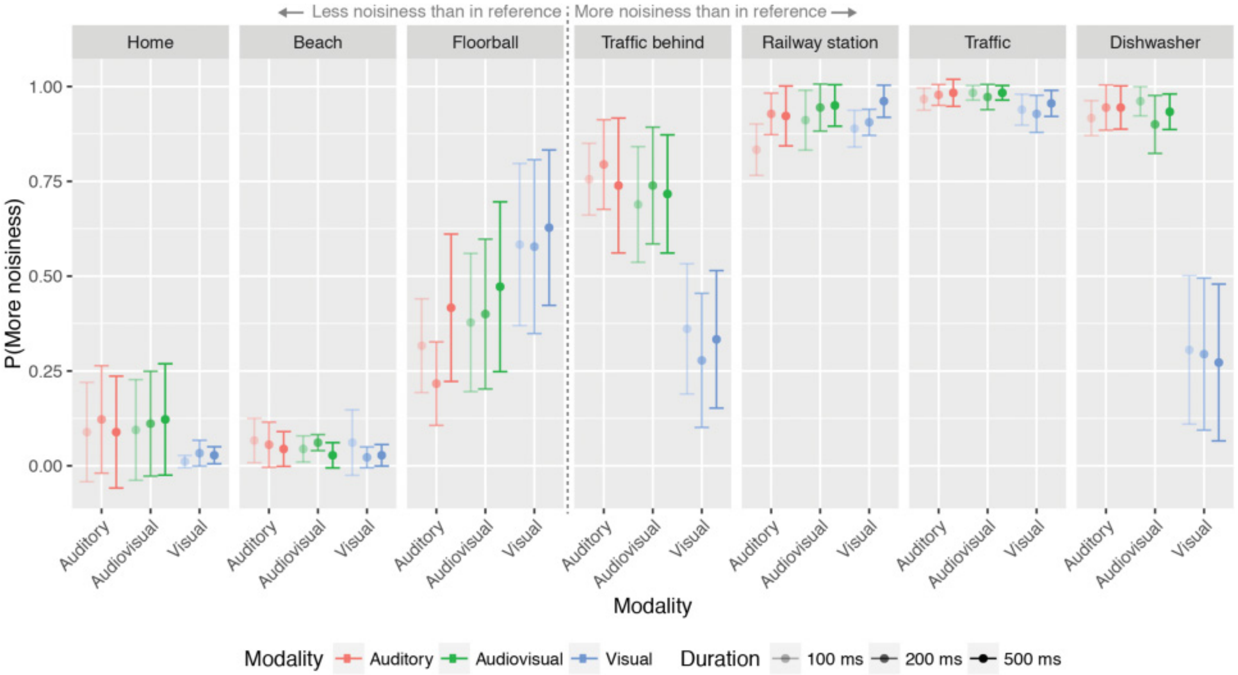

Fig. 8. Noisiness discrimination task between the stimulus scenes and the reference (#*Market square*). The amount of baseline noisiness is increasing from left to right; the reference scene is located between the 3rd and 4th scenes. Each sample mean results from 180 paired comparisons made by 15 participants evaluating the same pair 12 times each. The bars denote the 95% confidence intervals of the mean calculated from the 15 mean noisiness scores and assuming normal distribution.

Table VIII. Significant Main Effects in Noisiness Discrimination and Post-Hoc Tests with $p < 0.05$

Home:	-	-
Beach:	-	-
Floorball:	*Modality* $F_{(2,28)} = 5.81$, $p < 0.01$	V > A & AV
	Duration $F_{(2,28)} = 4.21$, $p = 0.04$	500ms > 200ms
Traffic behind:	*Modality* $F_{(2,28)} = 20.97$, $p < 0.01$	A & AV > V
Railway station:	*Duration* $F_{(2,28)} = 6.15$, $p = 0.01$	500ms > 100ms
Traffic:	*Modality* $F_{(2,28)} = 3.47$, $p = 0.05$	A & AV > V
Dishwasher:	*Modality* $F_{(2,28)} = 45.16$, $p < 0.01$	A & AV > V

labelled according to the early or late repetition, and its effect on the outcome is inspected. No significant main effects are found in any attribute.

4. DISCUSSION

The goal of this work was to study sensory contributions in natural scenes while performing a complex discrimination task, which is an advance from previous studies utilizing impoverished stimuli and simple tasks. We presented the participants with paired-comparison tasks (2-alternative forced choice) of brief natural scenes in three perceptual attributes: movement, openness, and noisiness. The scenes were reproduced in an immersive audiovisual environment, using either auditory, visual, or

Table IX. Effect of Repetition

Movement:	
Repetition	$F_{(1,14)} = 0.01$, $p = 0.92$
Repetition × *Scene*	$F_{(6,84)} = 0.16$, $p = 0.99$
Openness:	
Repetition	$F_{(1,13)} = 1.66$, $p = 0.22$
Repetition × *Scene*	$F_{(6,78)} = 0.43$, $p = 0.86$
Noisiness:	
Repetition	$F_{(1,14)} = 0.30$, $p = 0.59$
Repetition × *Scene*	$F_{(6,84)} = 2.90$, $p = 0.06$

audiovisual stimulation, and the task was to choose the scene having a higher level of the desired attribute.

The natural scenes were meaningful places that probably are easy to remember despite the very short stimulus duration. Using more abstract stimuli could reduce adaptation effects and make the task more dependent on sensory processing. However, we decided to maintain a high level of ecological validity and minimize the adaptation effects by randomization and rapid presentation of stimuli. In our data, we found no significant adaptation effects. Nevertheless, in future the number of stimulus scenes could be increased, while the number of repetitions could be decreased.

The chosen natural scene attributes have different perceptual characteristics: movement and noisiness occur in time dimension in both modalities, while openness is visually readily available and unchanged in time. In addition, the dominant sensory modality can be assumed to be visual for movement and openness, and auditory for noisiness. These assumptions are also supported by the majority of our results (mixed-model analysis in movement and openness, but not in noisiness). However, a few scene-specific cross-modality influences were found contradicting the simple sensory dominance explanation. Furthermore, given enough time, the auditory system was able to provide accurate information regarding not only noisiness, but also regarding the movement and openness of a scene.

The scenes and attributes that we used were identified as perceptually meaningful to humans in previous research; thus, we assume that our results will prove useful when creating convincing and natural virtual reality environments. Getting the first impression of a scene correct is a primary concern, as immersion is easily broken in the case of contradicting sensory stimulation. Our results highlight the importance of having the weaker modality supporting the dominating one in creating the situational awareness in bimodal conditions. On the other hand, we show the weaker modality to be sometimes able to distort the bimodal estimate (Scene #*Traffic* in openness), requiring careful planning of its usage.

Looking back at our hypotheses, the probability of successful scene discrimination was not better in the bimodal condition when compared to either of the unimodal conditions in any of the attributes. There were, however, scene-specific effects in which the modality that was assumed to be dominating performed worse than the assumed weaker modality that was found affecting the AV estimate instead. This was evident, but not significant, in movement in #*Beach* and #*Home* (Figure 6). In these scenes, the auditory scene seems to have a calming effect, which reduces the AV estimates of movement. Both scenes contain some movement that may not be obvious at first: the water in #*Beach* and the TV program in #*Home*. We hypothesize that the calm soundscape gives the perceiver a sense of awareness regarding one's surroundings and reduces the surprise effect of the visual movement, which, in turn, is amplified in the V condition. Understanding the interplay of visual and auditory factors in constructing a pleasant environment has been studied in the field of environmental psychology [Pheasant et al.

2010], but it should be taken more into consideration when creating models for visual attention or designing a soundscape for a specific virtual reality application, for example.

Similarly, in openness and noisiness, the bimodal estimate was never better than the best unimodal estimate in any of the scenes. In openness, an interesting cross-modality effect was observed in one scene, *#traffic*, in which the inclusion of A information reduced the audiovisual openness of the scene. The scene contained unpleasant and loud traffic noise, which potentially startled the participants and affected their perception of an attribute that was in other scenes significantly affected only by V information. This was the only significant case in which the stronger modality did not fully capture the bimodal estimate; thus, our hypothesis of the stronger modality not always capturing the bimodal estimate receives some support.

This finding has implications for the design of immersive audiovisual content. Poor planning of the use of spatialized audio may result in the unwanted side effect of altering the perceptual estimates of scene attributes from the expected. Recent findings of these effects are presented in Mendonça et al. [2015], in which natural 3D audio had a negative impact on a visually focused attention task.

Regarding our hypothesis that the unimodal conditions require longer stimulation to yield discrimination success comparable to the bimodal condition, we find support in the assumed weaker modality, especially in movement and openness. In movement, 500ms of auditory stimulation is required for the human auditory system to be able to extract enough information to discriminate movement in many scenes equally well as with 100ms audiovisual stimulation. There appears to be a reason for a well-spatialized sound field surrounding the perceiver in order to create an impression of moving sound sources. This discrimination may not be possible with auditory stimulation with less spatial information. The same effect was observed in openness in the outdoor scenes, in which the soundscape was clearly less reverberant than in the indoor reference scene *#Railway station*, but the perceptual evaluation of reverberance required more than 100ms to be accurate enough.

In noisiness, the assumed weaker V condition showed no improvement in the probability of success with longer stimulation; rather, the probability of success was either equally good as the AV and A cases or greatly impaired. In some scenes, such as *#railway station* and *#traffic*, containing a lot of movement and visual objects known to produce noisy sound (cars), a nonsignificant improvement of the AV condition over A or V alone was observed, pointing out the facilitatory role of the visual cues. Interestingly, visual information did not reduce the bimodal probability of success in noisiness discrimination when the visual scene appeared calm in contrast to noisy sound scenes (*#Dishwasher* and *#Traffic behind*). This is in contrast with the observations from the *#Traffic* scene in openness discrimination, in which the weaker modality was able to distort the bimodal discrimination.

In all attributes, the assumed stronger modality performed equally well with the bimodal case regardless of the stimulation duration, excluding the few exceptions mentioned earlier. This finding contradicts our hypothesis of the bimodal case yielding better discrimination results faster; rather, the bimodal performance follows the best unimodal case. In natural scenes, based on our results, the sensory weights often are very imbalanced, which results in sensory dominance-like effects, but does not exclude integration of unisensory estimates. Further studies are required to establish if these weights would change under different natural conditions (e.g., low luminosity), and whether the sensory dominance theory can be overruled in the case of multimodal natural scenes. Already, with the data that we have, the sensory dominance effect does not always seem to explain the integrated audiovisual discrimination accuracy in natural scenes. Rather, there are more factors at play. The sensory integration benefit would be the strongest in cases in which the visual and auditory signals were equally reliable.

It is important to note the limited number of scenes in our study when evaluating the generalizability of our results. The employed scenes represent typical urban environments with clear perceptual differences in scene movement, openness, and noisiness, as established in previous research

[Rummukainen et al. 2014]. However, there are other perceptual concepts at play in each scene as well (Table II). Therefore, making broad conclusions about audiovisual interactions in natural scenes is not viable based on our data. Our results should be viewed as further evidence of the value of multimodal input for accurate situational awareness and as a starting point for more detailed investigations into perceptual processes in the real world. We encourage the move toward more natural stimulation in perceptual studies and hope that the spherical videos, spatial audio, and data published along with this study help other researchers to begin experimenting with natural scenes.

Finally, multimodal perception of natural scenes appears to be beneficial for a human observer in attenuating or amplifying unimodal estimates of scene attributes in order to arrive at the best possible perceptual explanation of the characteristics of a scene. We were not able to show the bimodal discrimination to be consistently faster or more accurate than the best unimodal case. Rather, the bimodal estimates were robust against occasional misjudgments of any single modality (#*Beach* and #*Home* in movement, #*Traffic* in openness). Due to technological advances and the need for more lifelike virtual reality, there is an increasing demand to use multimodal stimulation in natural-scene perception studies to better comprehend the cross-modal processes affecting our understanding of the physical reality. The sensory integration mechanism has been hypothesized to perform in a near-optimal manner, taking into account the unimodal reliabilities, in previous psychophysical studies conducted with simple stimuli. However, most of our results can be explained by both the sensory dominance theory and sensory integration; further studies are needed to disentangle optimal integration from sensory dominance in natural scenes.

From a practical and applied point of view, we can state that, in immersive systems, each modality will be the strongest on different attributes. We must look into the multimodal effects attribute by attribute, which is different from how we have been evaluating the perception and quality of realistic presentations. Furthermore, the effect of the second modality is unpredictable. There is a tendency for the strongest cue to dominate, but which cue dominates varies and sometimes we find an effect of both cues. Today, common practice is to evaluate visual displays and spatial audio separately, which can lead to results that do not relate to reality.

5. CONCLUSIONS

We evaluated discrimination of real-world scenes in three perceptually meaningful attributes with short (100–500ms) unimodal and bimodal scene exposures. The scenes were reproduced in an immersive audiovisual environment with 3D sound and surrounding visuals. Movement and openness were found to be mainly visual attributes with some input from auditory information. In some scenes, the auditory system was found to be able to derive information about movement and openness that was comparable with an audiovisual condition already after 500ms stimulation. Noisiness was dominantly auditory, but visual information was found to be an aiding factor in some scenes. Cross-modality effects affecting global estimates of the scene attributes were found in movement and openness. In sum, we can assume that real-world scenes will be perceived differently depending on whether only one or two modalities are presented. Therefore, the perception and quality of a multimodal presentation should never be assessed by separating video and audio and looking into how they are perceived as self-contained entities.

REFERENCES

D. Alais and D. Burr. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology* 14, 257–262.

Alan Armstrong and Johann Issartel. 2014. Sensorimotor synchronization with audio-visual stimuli: Limited multisensory integration. *Experimental Brain Research* 232, 11.

D. Bates, M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1, 1–48.

Yi Chuan Chen and Charles Spence. 2010. When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition* 114, 389–404.

Yi-Chuan Chen, Su-Ling Yeh, and Charles Spence. 2011. Crossmodal constraints on human perceptual awareness: Auditory semantic modulation of binocular rivalry. *Frontiers in Psychology* 2, 1–13.

Ken Cheng, Sara J. Shettleworth, Janellen Huttenlocher, and John J. Rieser. 2007. Bayesian integration of spatial information. *Psychological Bulletin* 133, 4, 625–637.

Francis B. Colavita. 1974. Human sensory dominance. *Perception & Psychophysics* 16, 2, 409–412.

Verena Conrad, Mario Kleiner, Andreas Bartels, Jessica Hartcher O'Brien, Heinrich H. Bülthoff, and Uta Noppeney. 2013. Naturalistic stimulus structure determines the integration of audiovisual looming signals in binocular rivalry. *PLoS One* 8, 8, e70710.

Beatrice De Gelder and Paul Bertelson. 2003. Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences* 7, 10, 460–467.

Oliver Doehrmann and Marcus J. Naumer. 2008. Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Research* 1242, 136–150.

A. J. Ecker and L. M. Heller. 2005. Auditory-visual interactions in the perception of a ball's path. *Perception* 34, 1, 59–75.

Marc O. Ernst and Heinrich H. Bülthoff. 2004. Merging the senses into a robust percept. *Trends in Cognitive Sciences* 8, 4, 162–169.

K. K. Evans and Anne Treisman. 2010. Natural cross-modal mappings between visual and auditory features. *Journal of Vision* 10, 1, 1–12.

J. Gómez Bolaños and V. Pulkki. 2012. Immersive audiovisual environment with 3D audio playback. In *132nd Convention of the Audio Engineering Society*. Budapest, Hungary, 1–9.

David Hecht and Miriam Reiner. 2009. Sensory dominance in combinations of audio, visual and haptic stimuli. *Experimental Brain Research* 193, 2, 307–314.

A. Koene, Derek Arnold, and Alan Johnston. 2007. Bimodal sensory discrimination is finer than dual single modality discrimination. *Journal of Vision* 7, 11, 1–11.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.

Paul J. Laurienti, Robert A. Kraft, Joseph A. Maldjian, Jonathan H. Burdette, and Mark T. Wallace. 2004. Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research* 158, 4, 405–414.

J. López-Moliner and S. Soto-Faraco. 2007. Vision affects how fast we hear sounds move. *Journal of Vision* 7, 12, 1–7.

Catarina Mendonça, Olli Rummukainen, and Ville Pulkki. 2015. 3D sound can have a negative impact on the perception of visual content in audiovisual reproductions. In *21st International Conference on Auditory Display (ICAD'15)*. Graz, Austria.

Catarina Mendonça, Jorge A. Santos, and Joan López-Moliner. 2011. The benefit of multisensory integration with biological motion signals. *Experimental Brain Research* 213, 2–3, 185–192.

Alexis Pérez-Bellido, Salvador Soto-Faraco, and Joan López-Moliner. 2013. Sound-driven enhancement of vision: Disentangling detection-level from decision-level contributions. *Journal of Neurophysiology* 109, 4, 1065–1077.

Robert J. Pheasant, Mark N. Fisher, Greg R. Watts, David J. Whitaker, and Kirill V. Horoshenkov. 2010. The importance of auditory-visual interaction in the construction of tranquil space. *Journal of Environmental Psychology* 30, 4, 501–509.

A. Politis and V. Pulkki. 2011. Broadband analysis and synthesis for DirAC using A-format. In *131st Convention of the Audio Engineering Society*. New York, NY, 1–11.

V. Pulkki. 2007. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society* 55, 6, 503–516.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. http://www.R-project.org.

O. Rummukainen, J. Radun, T. Virtanen, and V. Pulkki. 2014. Categorization of natural dynamic audiovisual scenes. *PLoS One* 9, 5, e95848.

Charles Spence and Sarah Squire. 2003. Multisensory integration: Maintaining the perception of synchrony. *Current Biology* 13, 519–521.

Jye-sheng Tan and Su-ling Yeh. 2015. Audiovisual integration facilitates unconscious visual scene processing. *Journal of Experimental Psychology: Human Perception and Performance* 41, 5, 1325–1335.

Michel Treisman. 1998. Combining information: Probability summation and probability averaging in detection and discrimination. *Psychological Methods* 3, 2, 252–265.

J. Vilkamo, T. Lokki, and V. Pulkki. 2009. Directional audio coding: Virtual microphone-based synthesis and subjective evaluation. *Journal of the Audio Engineering Society* 57, 9, 709–724.

David R. Wozny, Ulrik R. Beierholm, and Ladan Shams. 2008. Human trimodal perception follows optimal statistical inference. *Journal of Vision* 8, 3, 1–11.

Shlomit Yuval-Greenberg and Leon Y. Deouell. 2009. The dog's meow: Asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research* 193, 4, 603–614.

Received October 2015; revised April 2016; accepted April 2016