Combining Texture-Derived Vibrotactile Feedback, Concatenative Synthesis and Photogrammetry for Virtual Reality Rendering

Magalhaes, Eduardo; Høeg, Emil Rosenlund; Bernardes, Gilberto; Bruun-Pedersen, Jon Ram; Serafin, Stefania; Nordahl, Rolf

Published in:
Proceedings of 2019 Sound and Music Computing Conference

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

COMBINING TEXTURE-DERIVED VIBROTACTILE FEEDBACK, CONCATENATIVE SYNTHESIS AND PHOTOGRAMMETRY FOR VIRTUAL REALITY RENDERING

Eduardo Magalhães
Universidade do Porto
eduardom@fe.up.pt

Emil Rosenlund Høeg
Aalborg University
erh@create.aau.dk

Gilberto Bernardes
Universidade do Porto
gba@fe.up.pt

Jon Ram Bruun-Pedersen
Aalborg University
jpe@create.aau.dk

Stefania Serafin
Aalborg University
sts@create.aau.dk

Rolf Nordahl
Aalborg University
rn@create.aau.dk

ABSTRACT

This paper describes a novel framework for real-time sonification of surface textures in virtual reality (VR), aimed towards realistically representing the experience of driving over a virtual surface. A combination of capturing techniques of real-world surfaces are used for mapping 3D geometry, texture maps or auditory attributes (aural and vibrotactile) feedback. For the sonification rendering, we propose the use of information from primarily graphical texture features, to define target units in concatenative sound synthesis. To foster models that go beyond current generation of simple sound textures (e.g., wind, rain, fire), towards highly "synchronized" and expressive scenarios, our contribution draws a framework for higher-level modeling of a bicycle's kinematic rolling on ground contact, with enhanced perceptual symbiosis between auditory, visual and vibrotactile stimuli. We scanned two surfaces represented as texture maps, consisting of different features, morphology and matching navigation. We define target trajectories in a 2-dimensional audio feature space, according to a temporal model and morphological attributes of the surfaces. This synthesis method serves two purposes: a real-time auditory feedback, and vibrotactile feedback induced through playing back the concatenated sound samples using a vibrotactile inducer speaker.

1. INTRODUCTION

Contact between interacting objects in a natural environment often conveys information about the objects themselves. For instance, when walking on a surface, the contact between our feet and the ground will provide auditory and haptic feedback, which in many cases may be sufficient for us to recognize the surface we are walking on [1]. For initial surface identification, humans may utilize vision [2], but studies have shown how the multimodal perception of surfaces contributes to our experience and recognition of surfaces [3]. If vision does not convey sufficient surface information, we may be able to discern it through other sensory channels. An example would be when gradually investigating a frozen lake, to test for safety. Simply looking at the lake will likely not provide sufficient information, so we gradually test from feedback; tapping or stepping on the lake to gauge its thickness from the multisensory feedback it causes. How does it sound? How does it feel? Does the stepping even produce visual feedback, from changes such as cracks or any other visible reactions to weight or impact? However, there seems to be no definitive consensus on what constitutes a texture [4], but perception of surface textures can be considered multidimensional (rough/smooth, fine/rough, slippery/resistant, etc.) as well as multisensory (experienced through haptics, vision and audition) [2]. As such, we expect surface interaction feedback, whether we need it to explore a surface (e.g. ice thickness) or simply expect it to be part of our contact experience in natural environments.

To our knowledge, most solutions to aural and vibrotactile feedback in virtual reality (VR) rely on static, fixed or synthetic solutions, triggered by binary states (i.e., on and off) or random variation. While these methods have been assessed as expressive and natural, in light of the degree they play within complex multimodal scenarios, no systematic evaluation has addressed the degrees of correspondence between modalities at finer temporal granularities [5]. In this context, our work strives for a method capable of generating audio streams with highly controllable nuances for aural and vibrotactile feedback using concatenative sound synthesis (CSS) [6]. Despite the lack of use-cases adopting this sample-based synthesis in VR, the technique has proven to be quite robust in generating dynamic, evolving and ever-changing sound textures from short audio excerpts. Additionally, it enables the creation of audio streams, at different temporal granularities.

The integration of CSS in VR can tackle two important limitations of the technique, as identified in [6]: 1) the evaluation of a descriptors' salience, notably the difference between the aural and the vibrotactile descriptor spaces, and 2) the definition of targets which convey both the finer degree of user-controllable actions interacting with the Virtual Environment (VE). Ultimately, we aim to foster a unified aural and vibrotactile framework, which stresses a

high level of control, adaptation and presumably a greater sense of immersion engendered by congruent sensory feedback.

In this paper, we advance a framework for the expressive sonification of a continous ground contact, applicable for driving type experiences (e.g. biking). More specifically, we aim towards driving experiences, where a part of the VR interaction includes forward-moving wheels, making contact with surface material, while moving across the virtual landscape. The paper presents a preliminary case scenario using an instrumented training bicycle, for interacting with two surface textures: asphalt and dirt road (covered with leaves). Using image texture features and their intrinsic structure derived from displacement maps, we aim to foster a method to dynamically map repetitive user actions to a CSS engine. In this context, the main novelty is the use of CSS for aural and vibrotactile feedback, which to the best of our knowledge, has not been thoroughly explored and tested. Ultimately, the contribution aims to explore a larger framework that addresses spatiotemporal rolling friction with a high level of visual-aural-haptic synchronism and congruence. We show some promising preliminary results, which propose that sound descriptor-based models within CSS, can provide a real-time adaptive method to induce vibrotactile feedback within VR.

2. RELATED WORK

The human somatosensory (haptic) system is commonly divided into cutaneous and kinesthetic senses. Cutaneous, also frequently referred to as tactile sensation, refers to exteroreceptors responding to stimuli across the entire surface of the skin, including touch, pressure, vibration, temperature and pain [7]. Kinesthetic sensation refers to the perceptual receptors in the joints, tendons and muscles which give information about position and movement of the body (proprioception). It constantly monitors if body movement is caused by self-directed motion or an external force (such as the vibration from the bike while crossing rough terrain) [8]. Haptic technology has the ability to evoke human somatosensory stimulation, by providing tactile or kinesthetic cues through e.g. ground-referenced haptic devices [9], for example, force feedback or vibration to physical steering props [10].

On a real bicycle ride, several sensory channels assist each other in maintaining balance and control, and vision is one of the most predominant modalities to elicit appropriate responses. However, other mechanical sensory systems are equally important to perform complex motor tasks. The vestibular system is responsible for sensing head movement, orientation and balance, and is tightly linked to the kinesthetic sensation and the visual system [7]. The multisensory integration of these different sensations assist humans in creating a unitary understanding of the world, e.g. through active haptic exploration. For example, the visual system can, with a greater range, anticipate incoming obstacles and changes in surface structure, which necessitates preparation of the body to respond accordingly, with both speed and precision to uphold balance and stability. Thus, recognition of surface textures and geometric variations is paramount to maintaining control of a real life bike. Equally, the perception of texture from a tactile perspective, can be viewed as a product of vibration during surface exploration with a lateral movement of the hand [11], or vibrotactile stimulation of the feet [3]. In VR, this sensation can be simulated through active haptics, such as vibrotactile actuators or force feedback systems [9]. It has been shown that mediating such information can increase the sense of realism of a virtual experience [12, 13], aid in the identification of surface information [14], as well as improving task performance and precision [15, 16].

Previous studies have shown that the human post-perceptual system will integrate conflicting information, in which case a bias towards the stronger modality may appear, also known as intersensory bias [17] or sensory dominance [2]. One of the most famous example of this, is the visuotactile cross-modal interaction of the rubber hand illusion [18]. In the experiment, a majority of subjects perceived a rubber hand as their own hand, when observing the rubber hand being brushed simultaneously as their own hand, while the real hand was in the same approximate position, but visually hidden. This sensory illusion persists, because the visual cues dominate the sense of proprioception. Furthermore, another study has shown that when exposed to a visuotactile discrepancy, the haptic sensation adapts to vision when visual stimuli is more reliable, but haptic exploration dominate vision when the reliability of the visual stimuli was decreased [19]. Similar results of haptic dominance was shown by [20] when participants were presented with ambiguous visual cues.

The perceptual system will go to great length to form and maintain a unitary experience. But while being an adept and capable system, it also shows a considerable reliance on specific pattern interpretations, and may easily be affected when its logic is challenged, whether intentionally or accidentally. Especially considering the latter; the range of multisensory feedback potential with immersive VR technology is comprehensive. The illusion of being present in the VE, requires that it represents a coherent perceptual experience with sensory consistency [21]. So while sensory integration from congruent stimuli may aid the intelligibility of an object or amplify the experience of its sensory feedback, incongruent stimuli may quickly become disruptive, confusing, or even hinder a sense of presence.

For VR experiences, where a central user experience (or interaction design aspects) depends on correct perceptual interpretations of VE content/objects (e.g. for interpretation of valid actions), consistent and congruent stimuli should be a crucial consideration [21]. Studies on natural interactive walking has shown how exactly auditory and active haptic feedback has been able to convey the multisensory experience of walking surfaces [3]. For an immersive experience of *driving* across simulated virtual surfaces, real-time auditory and vibrotactile feedback mediated with technology of high temporal and spatial resolution, is naturally likely to play an essential part as well, but differs from human plantigrade gait due to the mechanical differences between the interrupted surface contact of the

ISBN 978-84-09-08518-7 ISSN 2518-3672

feet and that of a spinning wheel which maintains constant and continuous contact.

While studies on exercise biking show that virtual nature environments can be used to augment the exercise experience for motivational purposes with elderly users [22], these augmentations focus primarily on visuals for environment mediation. Experiencing nature content and traveling the environments, showed to be defining parts of elderly users' motivation to exercise [23]. For some users, introducing a VR headset to increase system immersion [21] further improved motivation and the user experience, partly due to experiencing of increased presence, but more qualitatively from feeling 'closer' to the (virtual) nature environment [24].

2.1 Visual world capture

For realistic experiences of feedback, e.g. from interactive virtual objects, congruence logically depends on the perceived realism of objects themselves, through all sensory modalities. In a study on the perceptual relationship between audio, video and audiovisual quality, results showed that high image quality positively affects the perception of (accompanying) audio quality, and vice versa [25, 26]. As the quality of the visual display is steadily increasing, it can be cogently derived that rendering techniques for interfaces targeting other sensory modalities must be paid equal attention and enhancement, to maintain sensory consistency. Photo-realistic capturing techniques, such as photogrammetry, can be supplied by affordable camera hardware (e.g., a standard smartphone-embedded camera) with tangible post processing time, to enable accessible 3D 'scanning' procedures of real-life objects. These captures can be merged into 3D geometry meshes, or surface capture for texture mapping, etc. [27]. The process includes combining a vast amount of photographs of an object from different angles for perfect alignment and capture of depth, to create a mesh and textures that resembles the captured object. High resolution meshes can be simplified in several ways, but often includes creating normal maps to simulate the fine geometric surface detail, and thus achieving the same level of detail and perceived granularity, with a decreased amount of vertices. However, the issue with normal maps is that the meshes themselves are often flat, which will be revealed when perceived from extreme angles. Displacement mapping is a different technique used to render details on a simple mesh, but where surfaces are actually displaced, based on a greyscale texture map (also known as a displacement map). Displacement maps are derived from height fields and elevation is encoded in each texel in a range from 0 (black) to 255 (white), in an 8-bit image, generating an actual displacement of the vertices along the surface normals [28]. Black-colored pixels are translated to minimum height, and white-colored pixels are translated to maximum height. Displacement maps typically requires a high level of vertices to achieve a good result, however some displacement procedures also afford a tessellation phase (i.e. adding additional subdivisions to the mesh before applying the displacement itself) [29].

Fig. 1 shows the process of adding geometry to a mesh

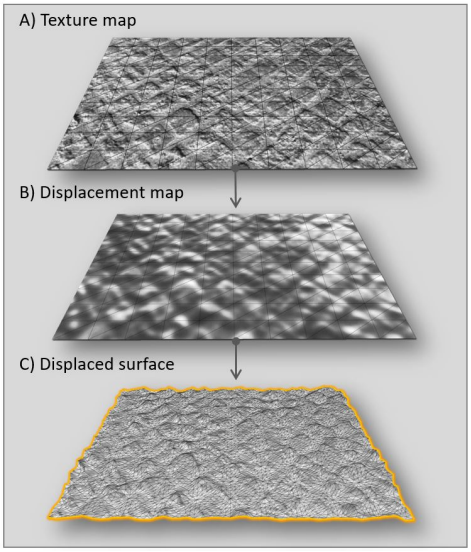

Figure 1. Texture mapping with an added displacement map. The results is a visually noticeable geometrical change of the surface, based fundamentally on the original texture map.

based on a displacement map and additional subdivision of the mesh (c). Here a texture map (a) is added to a primitive plane shape, and based on a displacement map (b) the mesh is tessellated and vertices are displaced along surface normals.

Although displacement mapping is a great way of achieving additional sense of depth and details, the addition of vertices also means an increased load on the graphical processing unit (GPU) [29]. For the developer, considerations towards optimization are always relevant, often related to the conflict of performance versus quality. For the purposes of this paper, the visual rendering of surfaces specifically (but not necessarily the remaining environment), should prioritize quality (displacement- over normal maps). Displacements and other geometric details in the visual surface representation, should be visually observable. Especially for moving forward, vision allows a user foresight of what surface features the vehicle will cross momentarily. It sets expectations for the translation of the visual cues into multimodal surface feedback. This can be the case for more prominent objects on the surface, but also simply for parts of the surface that are distinguishable or unique. And example would be a hard rock during on a soft dirt path, or a pile of crisp leaves on a sidewalk. The former (rock) would likely rely on displacement with multimodal feedback being focused quite a bit on the haptics, while the latter (leaves) would unlikely demand much displacement or haptic feedback, but would elicit strong auditory feedback.

2.2 Concatenative Sound Synthesis

Concatenative sound synthesis (CSS) is a sampling technique that creates audio streams by combining snippets from a large audio database [30]. It can be seen as an ex-

tension of granular synthesis [6,31], towards greater levels of control and automation by adopting content-based audio description methods from Music Information Retrieval (MIR).

Historically, CSS has been largely applied to the synthesis of sound textures/environmental sounds and music [32]. In the former category, of great relevance to the current work, CSS successfully tackles the pervasive post-production problem of extending a given audio clip, [33, 34], to more creative solutions within games, VR and interactive installations; and even to procedurally generate highly controlled nuances that match external actuators [31, 35].

CSS builds up a database of pre- or live-recorded *units* (i.e., segments or snippets with typical lengths of 50 ms to a few seconds) from an input *audio source*. Relevant sonic properties of each unit, such as pitch, loudness, noisiness, or spectral shape, are merged into a *feature vector*, which represents the units in the system. Due to their reduced dimensionality compared with the raw audio signal, feature vectors allow efficient search and retrieval from extensive data bases. New audio streams are created by specifying target queries, for which the best match is retrieved from the database to be played back.

The selection of attributes in the feature vector and target queries, as well as the metrics used to compare them are crucial to the system performance and quality. In this context, the nature of the input audio source, application domain and target definition (by navigating in a descriptor space) are fundamental to the parametrization of the system algorithms. For a comprehensive comparison of these variables and their implications in the musical results, please refer to [36].

3. METHODOLOGY

Fig. 2 shows the architecture of a concatenative sound synthesis engine for aural and vibrotactile feedback in VR. To the CSS prototypical component modules of this synthesis technique (in grey), we introduce a novel target definition method driven by texture map features from photogrammetric models of the provided surfaces (identified by their ID). Within this application context, particular emphasis is given to repetitive activities, such as walking, pedaling, and swimming, whose attributes are defined in the system by their angular velocity. Next, we detail each of the component modules of the architecture, using as a case-base scenario a bicycle ride on two surfaces asphalt and dirt road covered with leaves.

3.1 Source sounds

We recorded real bicycle rides on two different surfaces: asphalt and dirt road covered with leaves. The choice of these surfaces aims at designing a preliminary battery of multimodal tests, which enforce scenarios where the vibrotactile and aural feedback is known to have different impact. While the vibrotactile feedback of these two surfaces was expected to have a minimal discrepancy, the aural feedback was expected to be quite distinctive. At two

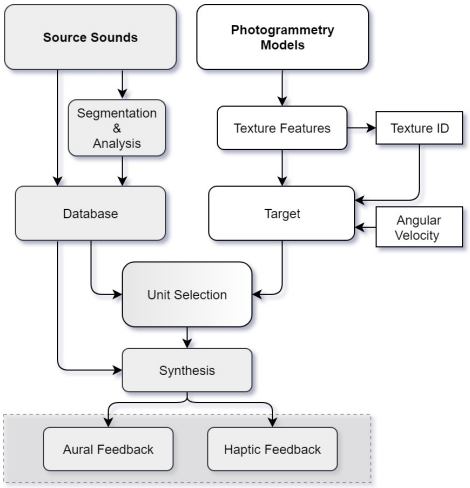

Figure 2. System architecture for a concatenative sound synthesis engine for vibrotactile and aural feedback (grey blocks) in VR. Target definition is driven by texture features from photogrammetric models, texture ID and angular velocity (white blocks).

of the recording sessions, microphones were mounted to a bicycle with 1) a *standard* cardioid microphone, pointing at the bicycle wheel to capture audio generated by the tire against the surfaces , and 2) a contact microphone attached to the chassis of the bike to capture the vibrations propagating through the wheel caused by the changing surface texture. The adopted audio capture techniques aimed at a clear distinction between synchronized auditory sources to test both the descriptors efficiency in discriminating the sources.

3.2 Segmentation and analysis

To segment the recordings, we chopped the signal into equal-length units of 50 ms. This simple and efficient segmentation method was favored instead of more complex and structure-aware segmentation methods (e.g. using peaks from a spectral novelty function), as the source sounds are largely monotonic and repetitive in nature. Furthermore, the adopted unit length typically ensures a high degree of stationary behavior across its duration, thus promoting a more robust analysis of the signal.

Each unit was then analyzed using the entire set of eight low-level audio descriptors within MUBU [37]: frequency, energy, periodicity, first-order autocorrelation coefficient (AC1), loudness, spectral centroid, spectral spread, and spectral skewness. Hence, an eight-dimensional feature vector was created for each unit. Each descriptor per unit is represented by the mean value of overlapping windows across its duration (window size $\approx 11.6ms$ with 50% overlap).

3.3 Database

A hierarchical database architecture was adopted to store all imported source audio files and generated data. The former is stored into a buffer and can be easily retrieved by

accessing chunks of audio by an index (or sample) range. The latter stores all generated data from the source files during segmentation and analysis in a database using a hierarchical structure. Its top-level hierarchy includes as many entries as the number of surfaces, identified by an ID. Within each surface ID, we can parse a sub-level with three fields: 1) unit number, 2) unit onset (in samples), and 3) descriptors (feature vector).

3.4 Target

Targets are defined from displacement maps information per texture, which is acquired from the surface shaders in the Unity3D game engine. Acquiring surface information is a process that requires several steps. First of all, to measure the speed of the bike, the angular velocity has to be converted to meters per second, and the position of the bike in world space (i.e the coordinate system of the game scene) has to be logged to figure out which mesh the bike is currently interacting with. In Unity, each game-object can be assigned a unique identification tag. Identification of surfaces is thus registered through a raycasting algorithm which originates from the approximated ground contact point of the front wheel of the virtual bike. When the raycast registers a collision with a ground-surface it identifies the surface-tag, and accesses the displacement map in the subshader, to read the corresponding normalized pixel value (between 0 and 1) in the texture coordinate (texcoord), defined by the position of the raycast-hit (see Fig.3.).

3.5 Unit selection

Units selection is the component module responsible for finding the best matching units to be synthesized. It aims to assess the database unit that best fits a particular target query–measured by its target cost, or distance in the descriptor space–but also by minimizing the spectral displacement between unit bounds–measured by concatenation cost. In this context, we focus mainly in the first metric, as the latter metric is implicitly modelled in the target definition. The better the targets capture the temporal nuances of the repetitive actions, the more unnecessary is the concatenation cost. Therefore, a simple real-time local search for the best matching unit is used without considering its surrounding temporal dimensions. Furthermore, we consider the adoption of jitter as a result of a flexible *k*-nearest neighbour with a user-definable *k* value, to be empirically tested.

3.6 Synthesis

Synthesis is done by two concatenating synthesis engines, which generate unique streams for each capture technique, i.e., aural and vibrotactile. These streams are then played back through their correspondent channel. The aural feedback is sent to the headphones and the vibrotactile feedback is sent to a low frequency audio transducer (ButtKicker BK-LFE).

Within each stream, selected units are concatenated with a Gaussian amplitude envelope and an 50% overlap. We adopt a spectral compressor-expander filter to enhance

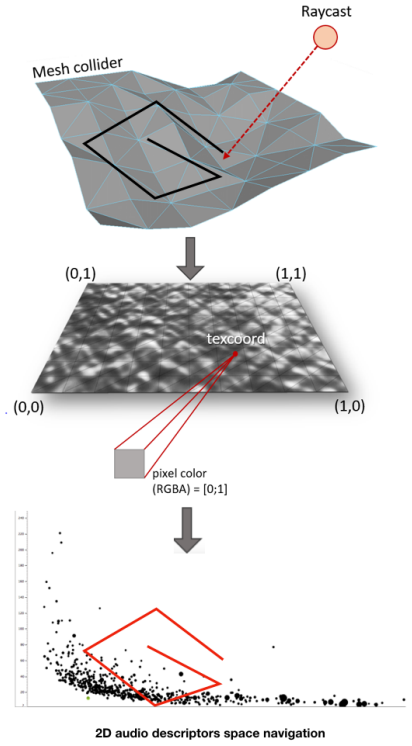

Figure 3. This figure visualizes the procedures of 1) acquiring surface information from a mesh through a raycast-hit, and 2) reading the corresponding texture-coordinate (texcoord) value in the displacement map, and 3) continuously sending that information to target units within the descriptors space once every frame.

the quality of the unit concatenation by minimizing discontinuities between the spectral peaks of adjacent units. Roughly, the signal process is done by applying on short-term windows ($\approx$ 50ms) a filtering mask resulting from the interpolation of the spectral content from neighbor windows.

4. PRELIMINARY EVALUATION

A preliminary evaluation of our framework aims to assess particular design choices of a unified CSS engine for both the aural and vibrotactile feedback in VR. In detail, we intend to promote a better understanding of the idiosyncrasies of the descriptor spaces resulting from our dual capture technique. Ultimately, we can endorse our informal perceptual hypothesis that the most expressive and salient features in aural and haptic descriptor spaces differ, and thus require different modelling strategies. To this end, we conducted statistical analysis to assess 1) which descriptors better represent each surface in a 2-D navigable space and 2) if the dual aural-haptic synthesis require different descriptors spaces to optimally navigate the corpus.

Following [34,38,39], we adopt a coefficient of variation, C_v, as a dimensionless (i.e. scale-invariant) measure of dispersion to identify the descriptors with greater salience.

As such, we aim to better discriminate the sound source in a 2-D navigable space and (conceivably) at a perceptual level. The coefficient of variation, C_v, is computed as the ratio of the standard deviation, σ, to the mean, μ, so that:

$$C_v = \frac{\sigma}{\mu} \quad . \tag{1}$$

The result expresses a percentage value of the descriptor extent of variability in relation to its mean.

Moreover, we adopt the Spearman's rank correlation ρ to determine the statistical dependence between the contact and standard (condenser) microphones on their descriptors degree of variability (or salience). The Spearman rank correlation is expressed by a value in the +1 to -1 range. +1 indicates a perfect association and -1 indicates a perfect negative association of ranks. The closer to 0, the weaker the association between the variables.

Our informal perceptual expectation is that the asphalt should provide a higher degree of descriptor dependency, given its low aural and haptic feedback, but may be more noticible in surfaces such as the dirt road with leaves, as its aural feedback is notoriously more prominent than the haptic feedback.

Table 1 shows the coefficient of variation, C_v, and mean values, μ, descriptor statistics for the two surfaces under study. The results are quite expressive across both surfaces and capture techniques for the spectral low-level timbral descriptors (centroid, spread, skewness and kurtosis). Conversely, the remaining temporal domain descriptors (frequency, energy, periodicity and first-order autocorrelation coefficient or AC1) show considerably less saliency. Moreover, the most salient descriptors (with the two high- est coefficient of variation, C_v, for asphalt are the same for both capture techniques (contact and condenser), while for the dirt road with leaves surface, the descriptors correspondence does not hold. To a certain extent, this reinforces our perceptual hypothesis that both haptic and aural realities are different and require different corpus and navigation models. The Spearman's rank correlation for each surface across the two capture techniques shows a high degree of dependency between the contact and standard (condenser) microphones ($\rho = .857$ with $p > 0.01$) for the asphalt surface. A weaker association is found for the dirt road with leaves ($\rho = .686$ with $p > 0.05$). These results align with our initial expectations that some surfaces will demand a greater degree of separation in the modelling of trajectories and descriptor space definition (to optimize the 2-D navigation).

Fig. 4 presents the 2-D navigation plots per audio capture technique, in which both surfaces are included. The graphical representation suggests that some level of overlap between surfaces can be adopted in the architecture and system design. In other words, we can envision a possible scenario where all surfaces coexist in the same corpus, without the need to specify in the target a surface ID to constrain the unit search.

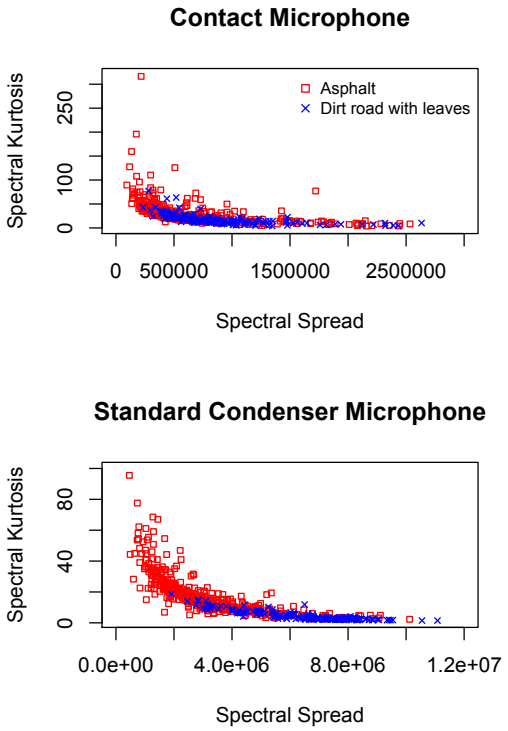

Figure 4. Two-dimensional visualization of the spectral spread and spectral kurtosis descriptor space for the two surfaces under study, i.e., asphalt and dirt road with leaves and for the two miking methods.

5. CONCLUSIONS AND FUTURE WORK

This paper presents preliminary work towards creating a novel approach to synthesize audiohaptic feedback from surface textures capture through e.g. photogrammetry. Besides facilitating visually realistic virtual objects, methods like photogrammetry represent new possibilities in a broader scheme. The ability to visually 'capture' reality into a realistic virtual representations, promotes methodological considerations for similar approaches to other modalities. The puzzle of constructing congruent integration of (synthetic) feedback, could become close to trivial, if feedback is constructed from capture of the actual visual, auditory and vibrotactile features of the real world object.

This work implementation and preliminary testing revealed that the combination of CSS with high realistic graphic assets, can provide a convincing and promising method to induce realism within immersive virtual environments. Furthermore, our preliminary results seem to indicate that using specific descriptors to analyze sound sources obtained from different capture methods can provide more precise clustering of sound units and sequences for each type of sensory feedback, including a further look into applications for realistic surface-haptic driving simulations.

We observed some limitations as well, at this point and related to some CSS intrinsic characteristics. A narrow

	Asphalt				Dirt with Leaves			
	Contact Microphone		Condenser Microphone		Contact Microphone		Condenser Microphone	
Audio Descriptor	C_v	μ	C_v	μ	C_v	μ	C_v	μ
Frequency	13	502.800	22	449.211	13	499.54	23	370.001
Energy	38	.005	32	.007	39	.009	34	.001
Periodicity	18	.371	41	.153	19	.388	53	.066
AC1	1	.967	1	.979	1	.963	2	.966
Loudness	7	-48.292	6	-47.194	7	-42.534	5	-40.083
Centroid	35	894.746	41	1294.465	30	1020.219	34	2846.533
Spread	**70**	692946.875	**57**	2942106.250	**50**	891232.511	27	6386084
Skewness	38	4.024	37	3.468	32	2.983	**72**	1.197
Kurtosis	**89**	31.758	**67**	20.014	**58**	17.701	**67**	4.681

Table 1. Coefficient of variation, C_v, and mean values, μ, statistics for the asphalt and dirt road covered with leaves surfaces per audio descriptor and for the two audio capture techniques.

sound corpus, if sound database is not broad enough. And lacking audio representations of specific interactions might result in less expressive and inaccurate sensory feedback.

This paper marks the first in a ongoing line of studies using the model proposed. Studies of interest include approaches to practical implementations of the feedback system (especially haptics, most likely), users perception of realism based on the multimodal surface feedback, possible implications for vection with VR bike augmentation, presence studies with elderly users, and a line of perception tests to further explore the best practices of the multisensory balancing between the modalities and techniques. Furthermore, future research should at least consider the impact on system performance when introducing shaders that are more demanding to the GPU, such as displacement mapping. However, alternative methods exists that do make use of the information derived from height fields to simulate displacement without adding additional geometry, e.g. parallax mapping. Such alternatives could still utilize the same methodology while being more affordable in terms of computational resources. A fundamental topic in the near future is to test the use of accelerometers as capture technique of the surface displacements. Our aim is to learn which capture technique better drives the generation of vibrotactile feedback using CSS, towards a definite methodology and a fully working prototype.

Acknowledgments

This work was supported by the Portuguese Foundation for Science and Technology (PD/BD/114140/2015).

6. REFERENCES

[1] Y. Visell, F. Fontana, B. L. Giordano, R. Nordahl, S. Serafin, and R. Bresin, "Sound design and perception in walking interactions," *International Journal of Human-Computer Studies*, vol. 67, no. 11, pp. 947–959, 2009.

[2] S. J. Lederman and R. L. Klatzky, "Multisensory texture perception," *The handbook of multisensory processes*, pp. 107–122, 2004.

[3] S. Serafin, L. Turchet, R. Nordahl, S. Dimitrov, A. Berrezag, and V. Hayward, "Identification of virtual grounds using virtual reality haptic shoes and sound synthesis," in *Proceedings of eurohaptics symposium on haptic and audio-visual stimuli: enhancing experiences and interaction*, 2010, pp. 61–70.

[4] R. L. Klatzky and S. J. Lederman, "Multisensory texture perception," in *Multisensory object perception in the primate brain*. Springer, 2010, pp. 211–230.

[5] A. Di Scipio, "Synthesis of environmental sound textures by iterated nonlinear functions," in *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, 1999, pp. 109–117.

[6] D. Schwarz, "Concatenative sound synthesis: The early years," *Journal of New Music Research*, vol. 35, no. 1, pp. 3–22, 2006.

[7] S. M. Breedlove and N. V. Watson, *Behavioral neuroscience*. Sinauer Associates, Incorporated, Publishers, 2017.

[8] M. S. Gazzaniga and T. F. Heatherton, *Psychological science: Mind, brain, and behavior*. WW Norton New York, 2003.

[9] J. J. LaViola Jr, E. Kruijff, R. P. McMahan, D. Bowman, and I. P. Poupyrev, *3D user interfaces: theory and practice*. Addison-Wesley Professional, 2017.

[10] D. C. Brogan, R. A. Metoyer, and J. K. Hodgins, "Dynamically simulated characters in virtual environments," *IEEE Computer Graphics and Applications*, vol. 18, no. 5, pp. 58–69, 1998.

[11] S. J. Lederman and R. L. Klatzky, "Hand movements: A window into haptic object recognition," *Cognitive psychology*, vol. 19, no. 3, pp. 342–368, 1987.

[12] L. Turchet, P. Burelli, and S. Serafin, "Haptic feedback for enhancing realism of walking simulations," *IEEE transactions on haptics*, vol. 6, no. 1, pp. 35–45, 2013.

[13] S. Lind, L. Thomsen, M. Egeberg, N. Nilsson, R. Nordahl, and S. Serafin, "Effects of vibrotactile stimulation during virtual sandboarding," in *2016 IEEE Virtual Reality (VR)*. IEEE, 2016, pp. 219–220.

[14] M. R. McGee, P. Gray, and S. Brewster, "Haptic perception of virtual roughness," in *CHI'01 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2001, pp. 155–156.

[15] I. Oakley, M. R. McGee, S. Brewster, and P. Gray, "Putting the feel in 'look and feel'," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 2000, pp. 415–422.

[16] V. Nitsch and B. Färber, "A meta-analysis of the effects of haptic interfaces on task performance with teleoperation systems," *IEEE transactions on haptics*, vol. 6, no. 4, pp. 387–398, 2013.

[17] R. B. Welch and D. H. Warren, "Immediate perceptual response to intersensory discrepancy." *Psychological bulletin*, vol. 88, no. 3, p. 638, 1980.

[18] M. Botvinick and J. Cohen, "Rubber hands 'feel'touch that eyes see," *Nature*, vol. 391, no. 6669, p. 756, 1998.

[19] J. Burge, A. R. Girshick, and M. S. Banks, "Visual–haptic adaptation is determined by relative reliability," *Journal of Neuroscience*, vol. 30, no. 22, pp. 7714–7721, 2010.

[20] M. O. Ernst, M. S. Banks, and H. H. Bülthoff, "Touch can change visual slant perception," *Nature neuroscience*, vol. 3, no. 1, p. 69, 2000.

[21] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 364, no. 1535, pp. 3549–3557, 2009.

[22] J. R. Bruun-Pedersen, S. Serafin, and L. B. Kofoed, "Restorative virtual environment design for augmenting nursing home rehabilitation," *Journal For Virtual Worlds Research*, vol. 9, no. 3, 2016.

[23] ——, "Motivating elderly to exercise-recreational virtual environment for indoor biking," in *Serious Games and Applications for Health (SeGAH), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–9.

[24] ——, "Going outside while staying inside—exercise motivation with immersive vs. non–immersive recreational virtual environment augmentation for older adult nursing home residents," in *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*. IEEE, 2016, pp. 216–226.

[25] A. Kohlrausch and S. van de Par, "Auditory-visual interaction: from fundamental research in cognitive psychology to (possible) applications," in *Human Vision and Electronic Imaging IV*, vol. 3644. International Society for Optics and Photonics, 1999, pp. 34–45.

[26] J. G. Beerends and F. E. De Caluwe, "The influence of video quality on perceived audio quality and vice versa," *Journal of the Audio Engineering Society*, vol. 47, no. 5, pp. 355–362, 1999.

[27] S. Lachambre, S. Lagarde, and C. Jover, "Unity photogrammetry workflow," 2017. [Online]. Available: https://unity3d.com/files/solutions/photogrammetry/Unity-Photogrammetry-Workflow_2017-07_v2.pdf

[28] R. Fernando and M. J. Kilgard, *The Cg Tutorial: The definitive guide to programmable real-time graphics*. Addison-Wesley Longman Publishing Co., Inc., 2003.

[29] J. Birn, *Digital lighting & rendering*. Pearson Education, 2014.

[30] M. Casey, "Soundspotting: A new kind of process?" in *The Oxford Handbook of Computer Music*, 2009.

[31] D. Schwarz and N. Schnell, "Descriptor-based sound texture sampling," in *Proceedings of the Sound and Music Computing Conference*, 2010, pp. 510–515.

[32] G. Bernardes and D. Cocharro, *Dynamic Music Generation, Audio Analysis-Synthesis Methods*. Cham: Springer International Publishing, In Press.

[33] M. Frojd and A. Horner, "Fast sound texture synthesis using overlap-add," in *International Computer Music Conference*, 2007.

[34] G. Bernardes, L. Aly, and M. E. Davies, "Seed: Resynthesizing environmental sounds from examples," in *Proceedings of the Sound and Music Computing Conference*, 2016.

[35] G. Bernardes, C. Guedes, and B. Pennycook, *EarGram: An Application for Interactive Exploration of Concatenative Sound Synthesis in Pure Data*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 110–129.

[36] N. M. Norowi, E. R. Miranda, and M. Hussin, "Parametric factors affecting concatenative sound synthesis," *Advanced Science Letters*, vol. 23, no. 6, pp. 5496–5500, 2017.

[37] N. Schnell, A. Röbel, D. Schwarz, G. Peeters, R. Borghesi *et al.*, "Mubu and friends–assembling tools for content based real-time interactive audio processing in max/msp," in *Proceedings of the International Computer Music Conference*, 2009.

[38] W. Brent, "A timbre analysis and classification toolkit for pure data," in *Proceedings of the International Ccomputer Music Conference*, 2010.

[39] L. Aly, R. Penha, and G. Bernardes, "Digit: A digital foley system to generate footstep sounds," in *Music Technology with Swing*, M. Aramaki, M. E. P. Davies, R. Kronland-Martinet, and S. Ystad, Eds. Cham: Springer International Publishing, 2018, pp. 429–441.