Medialogy AAU-CPH

Exploring Audiovisual Interaction through Rich Video

Master's Thesis
by

Kasper A. D. Boisen

As supervised by Stefania Serafin

[Intentionally left blank]

Preface:

Dear Reader.

This thesis was written in the period of February 2014 to May 2014, and I want to personally thank you for taking the time to read it.

The thesis features descriptions of several video productions in different variants produced for tests in the investigation. These videos are available on the included DVD for your interest and/or review.

Likewise, all illustrations used within the thesis are also available on the included DVD, as are the raw test results and answers.

Some tests in the thesis were conducted as an online questionnaire with videos embedded. This test is still available at the following link, should any interest arise:

http://testforme.redirectme.net

Special thanks goes out to:

Stefania Serafin, for making the thesis possible and for her helpful and concise supervision.

Mikael Peter Olsen, for continued support and friendship during writing and implementation, and for appearing in one of the videos created for the thesis.

Sebastian Hurup Bevensee, for the same support and friendship, and for featuring in two videos created in the thesis.

Without these people the thesis would never have been completed, and I am eternally grateful for all their help and support!

Contents

Motivation & Introduction

The academic world is full of outstanding research, theories and ideas with great potential – but despite results often showing promise for real world application, most are rarely developed beyond the theoretical stage. They often remain as clean and theoretical tests only proving that an effect occurs in controlled environments or in some cases in complete isolation from any and all other stimuli.

While theoretical and controlled experiments are the foundation of any idea and research, there is much to be learned from taking a theory outside controlled environments. How an idea is implemented into real-world scenarios, and how it fairs in disturbing and noisy environments?

These questions are the core of this investigation. Theory is very important, but so is implementation (!). A theory is just that, an immaterial result of thorough investigation – with its potential often unused. This is a central problem within the academic world, and it's somewhat unfortunate as there are undoubtedly great benefits in the effort of actually linking research to "the real world".

With personal interest in both video and audio, there are several established theories that I find potentially interesting if applied outside controlled and clean environments. One of these theories is called the "motion-bounce illusion", which states that sound alters visual motion perception (Sekuler R, 1997). This is the well-known experiment of two circles either passing through, or bouncing off each other if a sound is played as they intersect. The question is how this effect would work in a full, realistic video-environment. The initial experiments were basic and involving simple shapes, but one wonders if the same effect is present in a detailed, realistic and noisy video with distracting visuals and sounds.

If the motion-bounce experiment was to be tested with rich video material, the sound at the intersection should be equally rich, or some congruency between the two might be lost – a simple 'beep' might not suffice. The interesting thing is that there is no natural and preexisting harmony between image and sound (Chion, 1994). In essence, one has a lot of choices when combining images with audio (to video), and most combinations can be made to work, but with different expressions as the result. When it comes to creating realistic combinations of sound and image, one has to create a harmony in such a way that it the two seem inevitably connected even though they weren't in the first place (Chion, 1994).

This interconnection that is crafted between visuals and audio is what this investigation is about, and while the connections of visual and auditory stimulation are well documented, I see it as an interesting field to implement into finished rich environments and products.

In one area in particular is the audio-visual connection very important. With video elements receiving more and more post production, the areas of visual effects and sound effects are closely linked as everything in these areas needs to be created from scratch. When the designers create visual and auditory effects, the congruency between them is key to creating believable unified results.

Because of the importance observed here, and from personal interest, this study looked directly into how sound effects and visual effects affect each other as well as how their congruency matters. This was done with a basis in established audio-visual theories, which lead to several implementations into video examples of high production quality, as such attempting to apply theoretical experiments to 'real products'.

Pre-analysis & State of the Art

We constantly receive and absorb information through different sensory modalities. The properties of these stimuli are connected with physical laws so that auditory and visual stimuli caused by one event are linked together into a unified temporal perception (Joly, et al., 2001). For instance, we have little difficulty in linking the sound of one speaking to a specific person, even if there are multiple voices simultaneously (the cocktail party effect (Conway, et al., 2001)), indicating that the integration of information from separate sensory modalities is an important and effective part of our perceptual system.

While sensory modalities effectively link together in our perception of the world, they also interact, and potentially change each other. This is known to cause remarkable perception and identification of seemingly small details, as well as interesting illusions and changes within perception itself. As previously mentioned, it was of this investigation's interest to identify useful interaction within auditory and visual stimulation for application to video as rich media.

As such, the process of this study was to investigate audiovisual theories, implement suitable and successful theoretical aspects to rich media, resulting in an application onto video (see Figure 1)

Figure 1 - Process of the investigation

The investigation had a special interest in sound and its influence in video. It is commonly acknowledged that sound effects add value to film and video, as it aids the narrative, atmosphere and to accentuate moments. But the added value also works reciprocally. Sound alters the image and shows it differently than the image does alone (Chion, 1994). Likewise, the image makes us hear sound differently, and it is these effects that was of special interest to this investigation.

Sound helps unify the flow of images by bridging visual breaks with sound overlaps, which it also does by establishing atmosphere to the images, essentially establishing a framework that 'contains' the image. It also creates a unity with music that exists nondiegetically, further tying images together (Chion, 1994). This unity was the strongest argument for pairing sound and image, and while it is not easily achieved successfully, it is both important and powerful in the creation of video which created the foundation for the investigation's main focus of audiovisual interaction.

Audio-Visual Interaction

The interaction between sensorial modalities often leads to a change in perception. One modality may change the perception of another as the two are fuse together into a single perceived event. As such, *the interaction between two sensory modalities can be defined as the alteration of a certain percept in one sensory modality due to the presence of a stimulus in another sensory modality* (Para, 1999).

A prime example of this audio-visual interaction is the so-called "motion-bounce" illusion, in which two identical disks move to a point of intersection, where they are perceived to either 1) pass through each other, or 2) bounce off each other if accompanied by a sound at the intersection. The animation is of course the same, but the perception changes depending on auditory stimulus. Without the sound, only 20% see the disks as bouncing off each other, while 60% report the bounce with the sound present (Sekuler R, 1997), making this a clear example of an interaction between the auditory and visual modality. This effect can also be attenuated if the experiment is combined with what is known as "auditory flankers" just before the intersection (Shimojo, 2001). For this investigation, the interest was if this effect only occurred with simple circles in the clean environment (as was the case with most previous experiments) or if it was also present with detailed visuals and realistic sounds.

In general, it is stated that a valuable way of examining multisensory interaction is by creating stimulus of conflicting information in different modalities. (Para, 1999) Here the conflicting information often doesn't segregate into isolated percepts for each modality, and they remain as a unified perception of the event.

One consequence of this mismatch of sensorial stimuli of interest to this project was the fact that visual stimulus can influence the perceived distance of an auditory stimulus (J. M. Brown, 1998). In experiments it has been proven that an auditory stimulus can be perceived as in, or out of range when also presented with a (silent) visual object being in or out of range too. It is possible that this effect is also present in video, and it might also potentially work in reverse (audio changing the perceived distance of a visual stimulus).

Dominance of one modality over another is agued to be done on a basis of which is more suitable for a task (modality appropriateness hypothesis). Vision is regarded as the generator of the concept of space while audition is considered as acting as timekeeper (Hirsh, 1972). This is supported by the fact that visual flicker is undetectable after 50-100hz, where auditory modulation is detectable up to around 400hz. It is considered likely that a sound effect can make a visual effect or element seem faster if dominant in a scene of both stimuli.

Interestingly, there is also an audio-visual interaction dealing with the perceived quality of video material affected by sound. This particular effect deals with how the quality of audio material affects the perception of visual material, essentially making the sound in video the dominant stimuli with regards to perceived quality (Joly, et al., 2001). For this study, this effect meant that any lack in visual quality, could be abated with high sound quality, an important note for development and implementation. This also meant that any presence of poor sound quality in the investigation's products could negatively affect the visual quality, which should be avoided if possible.

Attention to a specific stimulus can be impaired by stimuli of other sensory modalities. This can be viewed as a problem and was one of the issues this investigation faced. Taking theory out of controlled environments means that there was a greater risk of overriding stimuli, rendering experiments and effects false. It can however also be used as a tool, with which one can direct attention directly and further emphasize aspects of experiments, if designed correctly. For implementation and design of experiments and, the audio and visual elements need to be evaluated both separately and together because of the potential interaction between them. This interaction is what this investigation pursues, requiring the tests to be designed so that it can be measured.

Designing Sound

With sound design, one creates the audibility of a world and its events. Sound is carries information about the world, and when you listen to it, a communication takes place. To a certain degree, sound designers are engineers of communication, as any acoustic event can exist as a sign carrier that communicates information about the world.

In order to create successful sound products, the designer needs to make careful decisions on the basis on how the listener perceive a sound and what communication takes place during a sound event. For this reason, sound design requires an understanding in semiotics to correctly use and apply this communication for its intended effects (Jekosch, 2005).

Within semiotics, a sign is a mental unit standing for- or pointing at something else than itself. This includes examples such as traffic-lights, flags and speech in general. For these, it is necessary to know the relation between the primary objects perceived and the concept they are denoting. Associating and object with the sign it carries, is referred to as "semiosis".

The job of a sound designer is also to create the reality of specific locations, but most importantly it deals with the expression of audible events. "*Within every character, object, and action on screen there can be generated a potential sound that may give further dramatic impact to the scene and story*" (Sonnenschein, 2001). The challenge is when to reveal this drama and when to leave it hidden. How this is to be revealed is a "sonic coloring" meaning that it can be designed in many different ways, all having different impact on the event. A general rule when intending to create drama with sound, is to utilize the absence of sound, so that a select few sounds will register more effectively with the audience creating a high aural contrast. (Sonnenschein, 2001). Another efficient way is to manipulate the distance or intimacy of a sound, as close-sounding events are generally perceived to be more dramatic and important.

Describing sound has always been difficult. While technical properties can take a description far, subjective explanations are often necessary to convey an audio element properly. Following is a list of subjective bipolar extremes of sound effects that are used in the investigation, derived from the work of David Sonnenschein.:

• closed-open
• loud-soft
• dry-echo
• low pitch-high pitch
• near-far
• empty-full
• harmony-dissonance
• friendly-menacing

These subjective descriptions will of course be accompanied by the precise terms of frequency, amplitude, attack, decay, etc.

Sound itself implies agitation or displacement. If something is making a sound it is moving in one way or another, with very few exceptions (Chion, 1994). Because of this, we automatically link sound to moving objects, whereas the lack of any implies a static environment. This is also why sound is an important tool when giving life and movement to synthetic elements, helping a successful integration of sound and visuals elements.

Cross-modal Sensory Interaction

As mentioned, our senses interact with each other, and change our perception. The human sensorimotor system experiences the world as a whole by synthesizing and merging stimuli from multiple modalities. Our perception attempts this merging at all times, in some cases leading to illusions and even unwanted and uncomfortable effects such as motion sickness/simulation sickness, being the result of mismatching stimuli (Biocca, et al., 2001). While this investigation was unlikely to cause any such issues, the crossmodal interaction of the senses was of interest with its potential implementation into rich media, in which one can purposefully manipulate the congruency between stimuli, or seek to minimize errors between them for effect. Achieving this congruency is a pursuit of intermodal integration and a unity of stimuli.

This intermodal integration can be used to have one modality "fill in the blanks" of another, with information from one stimuli disambiguating input from another drawing attention to objects or events (Biocca, et al., 2001). Intersensory interactions does vary, and some claim that "*the visual sensory channel is more likely to skew the interpretation of information processed by the other senses.* " (Welsh & Warren, 1980). While this isn't a recent claim, this investigation would like to suggest that the auditory sensory channel is also likely to skew visual interpretations in some cases, which is being followed up later in the thesis.

Cross-modal sensory interaction has also been categorized by Biocca et al for Virtual Environments, and this classification was also deemed suitable for this study's focus on video. In particular, the categories "intersensory Biases and adaptations", "Cross-modal enhancement or modification", and "Cross-modal transfers or illusions" (Biocca, et al., 2001).

- intersensory Biases and adaptation:

Stimuli from at least two sensory modalities creating conflicting information about a virtual object, for instance its spatial location. As a person attempts to integrate this discrepant information they may experience sensimotory illusions biased towards a sensory modality. This can for instance cause incorrect perception of distance.

- Cross-modal enhancement or modification:

Stimuli from a sensory channel altering or enhancing the perception and interpretation of stimuli from another sensory channel. This interaction may result in changes of detectability, perceived intensity, and perceived fidelity. An examples includes an increased perceived visual fidelity following an increase of auditory fidelity.

- Cross-modal transfers or illusions:

The extreme of crossmodal sensory interaction. Stimulation of one sensory channel leads to an illusion of stimulation in another channel, not actively affected by any connected stimuli. This level is effectively synesthesia, seeing colored or pulsating lights when hearing sounds and the like.

'Intersensory biases and adaptation' together with 'cross-modal enhancement or modification' are of special interest as both hold some promise of improving or changing perception, potentially useful in creating interesting and/or beneficial effects within the production of video, which is what this investigation examines.

Initial Success Criteria

In essence, this investigation was be based upon audio-visual interaction. Drawing on inspiration from Michel Chion, it aims to do the opposite from his research. Where he draws theory from film and video, this investigation draws video from theory.

Audio-visual illusions was studied and taken from theory and applied to rich media. As such, the first success criteria was to have these illusions functional in any final products regardless of noise or interfering elements. This was not easily achieved, but having this difficult goal and success criteria was intentional.

As an elaboration on the previous criteria, the investigation also aimed to successfully invoke crossmodal sensory interaction – which itself is a requirement for many audio-visual illusions.

All video material as product for testing must attain a level of high production quality. This goes for both visuals and audio as this study was applying theory to rich media, thus necessitating high quality.

To Summarize. The initial success criteria of this study were as follows:

- Successfully implement simple audio-visual theories to rich media.
- Achieve products of high production quality for testing.
- Invoke crossmodal sensorial interaction with rich video.

As such, the intents of this investigation was expressed through the following problem statement:

To what extend can crossmodal sensorial interaction be achieved through rich video of high production quality?

Investigation and Experiments

Due to the nature of this investigation, the thesis was structured around 5 different experiments, each with its own theory, often building upon previous knowledge and results. This resembles iterative development, but does not aim to produce a singular fully developed product – instead it investigates several related areas that build upon each other.

Overview of Experiments:

To ensure an overview of this approach, the following briefly describes each experiment and how they build upon each other:

- **Experiment 1: Audiovisual perception of Distance**
  An initial investigation of the audiovisual perception of distance. This experiment displays conflicting audio visual information and attempts to determine dominant stimuli in 2 levels of rich media as video, dealing with cross-modal interaction of intersensory biases and adaptation.

- **Experiment 2: Manipulating perception of direction and trajectory**
  This experiment builds on the perception of distance, and gives synthesized graphics and sound a velocity and trajectory in an augmented video. Dominant stimuli was again sought out, but this time with respect to a moving audiovisual object with incongruent stimuli, which also dealt with intersensory biases and adaptation.

- **Experiment 3: The motion bounce illusion in rich media**
  The motion bounce illusion is often mentioned as an example of audiovisual interaction. With the original experiments always being shown as abstract figures intersecting or not, this experiment attempts to take the experiment into realistic video to examine if it still occurs. The experiment deals with cross-modal enhancement or modification as sensory interaction.

- **Experiment 4: Expectation following obscured audio-visual event.**
  During the second experiment a question was raised about our expectation following audio-visual stimuli in video. As the second experiment cut to black before an event, participants were asked what they expected would happen, which gave rise to the question of how much

auditory stimuli could change an expected outcome of identical visual elements in video. The sensory interaction within is that of cross-modal enhancement or modification.

- **Experiment 5: Combining knowledge and elements from previous experiments.**
  A combination of selected elements and knowledge from the previous four experiments, joined into a small video narrative of the highest possible production quality. Due to the uncommon events described in these previous experiments, combining them lead to a highly augmented video sequence which was tested for realism and quality.

Experiment 1: Audiovisual Perception of Distance

As previously mentioned, there was an initial interest in the effects of established theory in rich media and environments. The first and preliminary interest was in the perception of distance through cross-modal stimulation. The question posed was if there was a dominant stimuli in a rich environment, and if there is an actual difference between rich and simpler environments, with regards to the perception of distance.

The initial experiment aims to present conflicting information through audio-visual stimuli, asking participants to judge the distance to the source of a sound and visual effect. This experiment followed the investigation's desire to produce video of high quality, with its visual effects and sound effects created on the basis of established theory. This gave the experiment both options and limitations to the product intended for testing.

Keeping the production as photo-realistic as possible, the visual effects and sound effects was 'simple' for the sake of the early test. With this criteria and the desire to test distance, the test was designed to work with a lightning strike in a harbor area readily available for testing. The intention was to present a lightning strike that was 1) close, or 2) far away, and have pre-recorded lightning-strikes or thunder from different distances play at both. This meant that a close-up lightning strike was tested with both a distant sound effect and one close-up, to see if the auditory stimuli could overwrite the visual stimuli, with the same for the strike far away. Also of interest was the level of detail and/or noise in the picture, and if this influenced this incongruent presentation of stimuli. For this, the background was altered to be either simple or rich with objects, as well as added rain in the rich version. The test was constructed as a video with real footage, synthesized visual effects and pre-recorded sound effects with manipulation. Furthermore, the simplification of the rich unaltered footage was done with compositing, adding simplified elements from multiple sources.

Following the classification mentioned earlier (Biocca, et al., 2001), this test deals with intersensory bias and adaptation as a crossmodal sensory interaction. This experiment exists in this category because of its manipulation of incongruent stimuli, and it should be possible to identify dominant stimuli as conclusions.

Perceiving distance with sound

To test the perception of distance through sound, auditory depth cues had to be selected for manipulation. The simplest of which was intensity. To be elaborate, this cue deals with a sounds relative intensity to its surroundings, where the intensity decreases over distance following the inverse-square law (Wolfe, et al., 2009). If a sound occurs in a free field (without reflections), radiating into space uniformly in all directions, the sound pressure level, and thereby intensity, will decrease by 6dB for each time the distance from the source is doubled (Labor, 2014). See illustration below as example (Figure 2):

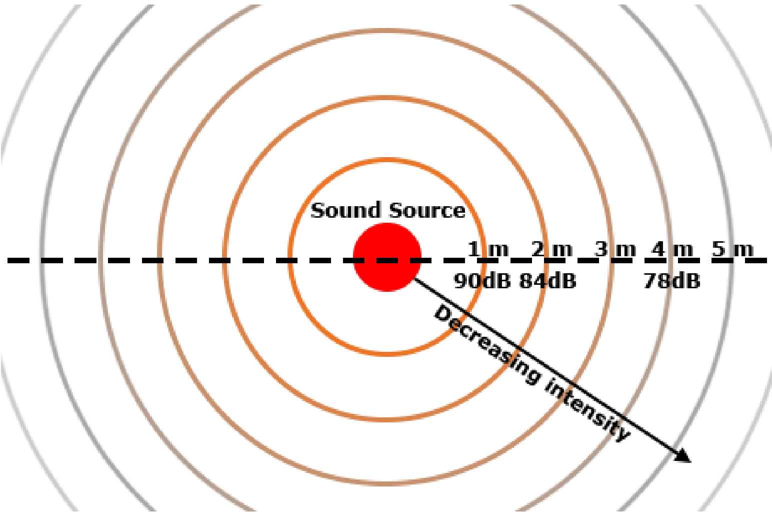


Figure 2 - Decreasing intensity over distance

Another auditory distance cue is that of spectral composition. This is referring to the result of the air quickly absorbing high-frequency sounds which is most significantly audible over distances further than 1000m (Wolfe, et al., 2009). While the higher frequencies decrease rapidly in intensity over distance, the low frequencies decrease slowly, as often heard in thunder cracks far away. This gives the experiment yet another tool to display distance with audio – effectively implemented with low-pass filters.

While distance to a sound is most accurate within 1 meter of the listener, there is rarely sound in rich media being portrayed as within this short distance. At longer distances, our accuracy decreases, and we often tend to underestimate distance determined by remote auditory stimuli (Zahorik, 2002). This inaccuracy, was an argument for having a sound source in the experiment far away as well as up close, to see its effects in rich media. As with other stimuli, this perceived distance works best if the source is moving – which was explored in the second experiment. Studies have also shown that the brain effectively perceives sound velocity when integrating audiovisual information and that it does this over a wide range of temporal gaps between auditory and visual elements (Sugita & Suzuki, 2003).

Experiment 1 Overview

Testing our perception of distance with manipulated auditory depth cues led to the creation of 2 videos, each with a different version of audio. As the experiment was also interested in effects of rich/noisy backgrounds and depth perception, two versions of this was also created each with a different background resulting in 4 videos in total – 2 test, and 2 control:

- The test consisted of two sets of visual backgrounds; Experiment 1a & 1b.
    - With and without visual clutter (noise) resulting in rich and simple backgrounds.
- Two lightning strikes occurred in each video, one at medium distance and one at long distance.
- The sound was created in two versions for each set of visuals
    - One with the sound of lightning swapped, ie. Medium-distance has long distance sound.
    - And one control, with matching sound.

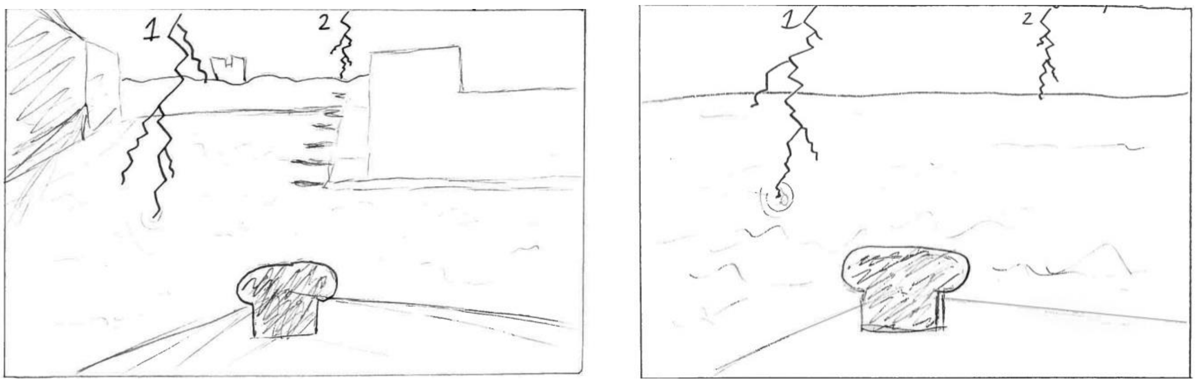


Figure 3 – Early sketch of Experiment 1a (left) and 1b (right).

The video with added noise and background was created as standard with the simple version being created by removing elements from the first. The video with rich background featured footage of a harbor-area with buildings at different distances and of different sizes, together with some post-production rain as both visual and auditory noise.

The test was primarily focused on investigating if there was a dominant stimuli in the perception of distance to audiovisual events in video. Additionally the test was also investigating the following hypothesis

> **H_1:** There is a significant difference in perception of distance in audiovisual with rich or simple backgrounds.

Participants were asked to pick which lightning strike was closest after watching a video with the two strikes. Their choice determined if they perceived the distance by one or the other stimulus. Furthermore, participants were asked to elaborate on their choice.

Design & Implementation

There exist a decent amount of research in lightning and thunder phenomenon, but it was difficult to come by any dealing with visuals of lightning in video. Because of this the design of the lightning element for the test-video was based on a personal case study of video and images of lightning.

The first observation was that the glow of lightning often has a color, and is not just white and bright. This color ranges from blue, through purple and violet to almost red and orange, suggesting that different recording methods produce different colors when subjected to the bright and sudden light of lightning. One assume white-balance affects this color. See Figure 4 for observations during one study.

Figure 4 - Lightning Strike: Visual elements. Based on image from: Fir0002, January 2007

Further observation established that lightning caught on video flickers in intensity in a period from 0.1 to 1 second. This flickering is induced by the frame rate of the camera from the much faster flickering of the real strike cause by current travelling in periods up and down the core in strikes and return strikes. Lastly, a yellow-orange tint often appears on the core itself close to the point of impact of a lightning strike caught up close on video – the cause of this was unknown, but was later added to the implementation as it appeared in several photographs and video.

Lightning strikes at medium-far distances are often partly occluded by clouds, also enhancing the fact that they originate *in* the clouds. Furthermore, the core and lightning strike in general has a strong glow or aura element lighting up water vapor in the vicinity forming a gradient from the core and out. It is this glow that is often in various colors in daylight photography and video.

For further reference, a video of a close-up lightning strike was analyzed for use during implementation (Figure 5).

Figure 5 – Still image from video of a Lightning Strike in the Bay of Kotor in Montenegro

This video shows a lightning strike off the coast hitting the water which was also the intention of the test. It was filmed with a camcorder at 30 frames per second (29.970 @ NTSC), and the lightning strike was visible for approximately 21 frames of this video being 2/3s of a second. This duration was used as benchmark for the creating of the test video, which was created in 25 frames per second to European and progressive field standards. Converting from NTSC to PAL, the strike should last for around 17 frames at the new frame rate for the test.

The video also doesn't show any secondary forks from the main core of the strike. This was because of the proximity of the strike, and the fact that it makes contact with the ground. The strike was blue to violet tinted and with a hint of orange at the point of impact on the water surface. This was all simulated during creation of the lightning strikes for the video.

The core pulses in thickness for the duration but was also obfuscated by artifacts appearing because of the camera's rolling shutter, often hiding the lightning or overexposing it due to the frequency mismatch of the lightning and the shutter. This was only partly simulated for later implementation.

As previously described, the sound of the lightning strike vary over distance. The sounds were chosen from stock archives for their realism, on the basis of 3 aspects; Attack, reverberation, and frequency range. For the near strike sound, the sound needed a short-immediate attack, some reverberation and several high-frequency elements. For the strike further away, it needed a smoother attack, with a lot of reverberation with some, but few high-frequency elements, and dominant low-frequencies.

With the test examining stimuli and distance perception, the test videos were constructed in such a way that none of the lightning strikes were at any extreme distances (neither very close nor very far). The aim was to have slight level of ambiguity to all aspects while still keeping the distances distinguishable. Both the visuals and sound for the lightning strikes was at medium to far distances from the camera, with the conflicting distance cues being almost plausible for each other. The test thus allows for one or the other stimuli to appear dominant, with the participant to determining the closest strike with one or the other stimuli.

The lightning strike was composited in Adobe After Effects. A shot of the harbor area was used for both the simple and rich video, wherein the simple had the buildings "painted over". This proved more difficult than first anticipated because of the reflections of the buildings in the water. These reflections had to be manipulated so that they didn't look like buildings, as the water itself had to be kept for the video, and the reflections were instead made to look as if reflecting darker clouds, which were also added during compositing.

The lightning-strike was created with a specialized plugin called 'CC Advanced Lightning', in which parameters were tweaked to make it look closer or further away. These parameters included forks and thickness of the core, along with turbulence and complexity of the path created on the basis of fractals. Glow was added to the core, as were a few extra touches to the impact on the water, such as illumination and extra intensity. Reflection of the strike in the water was also added, along with occluding clouds for the strike furthest away. A slight blue-purple hue was added as this was observed in video case studies.

All of the above had to be animated in intensity and with simulated overexposure to correctly mimic the flickering of lightning. As mentioned, this flickering lasted for 17frames (@25 FPS), accompanied by an illumination of the surrounding buildings in the rich example.

The rich and simple video were identical apart from the surroundings and an added simulation of light rain in the rich video. The lightning strikes were in the same places, albeit one occluded by buildings in the rich version. The two strikes were made to be at distinguishable different distances, but ambiguous enough to be potentially be affected by auditory depth cues.

Each had a quiet soundscape of a harbor front, one with added light rain. The lightning strikes were at high and medium amplitude, and with more or fewer high frequencies respectively, to simulate their distance. The recordings of the lightning were from the 'BBC Complete Sound Effects Library' and from the video game 'Left 4 Dead', and were specifically selected to meet the criteria of frequencies and timbre. The sound of a lightning strike is the extreme end of full, and menacing sound and is very loud compared to other sounds in the harbor soundscape.

The test was to implement the rule of the inverse-square law on the loudness of the thunder-strike (120dB (Vavrek, et al., 2014)), applying the correct lowering of intensity according to distance (see Figure 2) and with the application of low-pass filter to further emulate distance through spectral composition. The inverse-square law was however only partly applicable, as it depended on selected intensity of the audio equipment during playback.

The amplitude, and delay of the lightning strike sound effects were as such calculated on the basis of their distance to the camera. With sound traveling at 340m/s, the delay was easily calculated. 0.9s for the close strike, and 2.2s for the far strike (see Figure 6). The amplitude of a lightning strike in real life follows the inverse-square law as previously mentioned, but some adaptations had to be made for the implementation into video (derived from the allowed reasonable amplitude of audio equipment and the actual amplitude of a lightning strike). A decision was made to have the far lightning strike to have at least 50% amplitude, meaning that it was made to be at least -6db quieter.

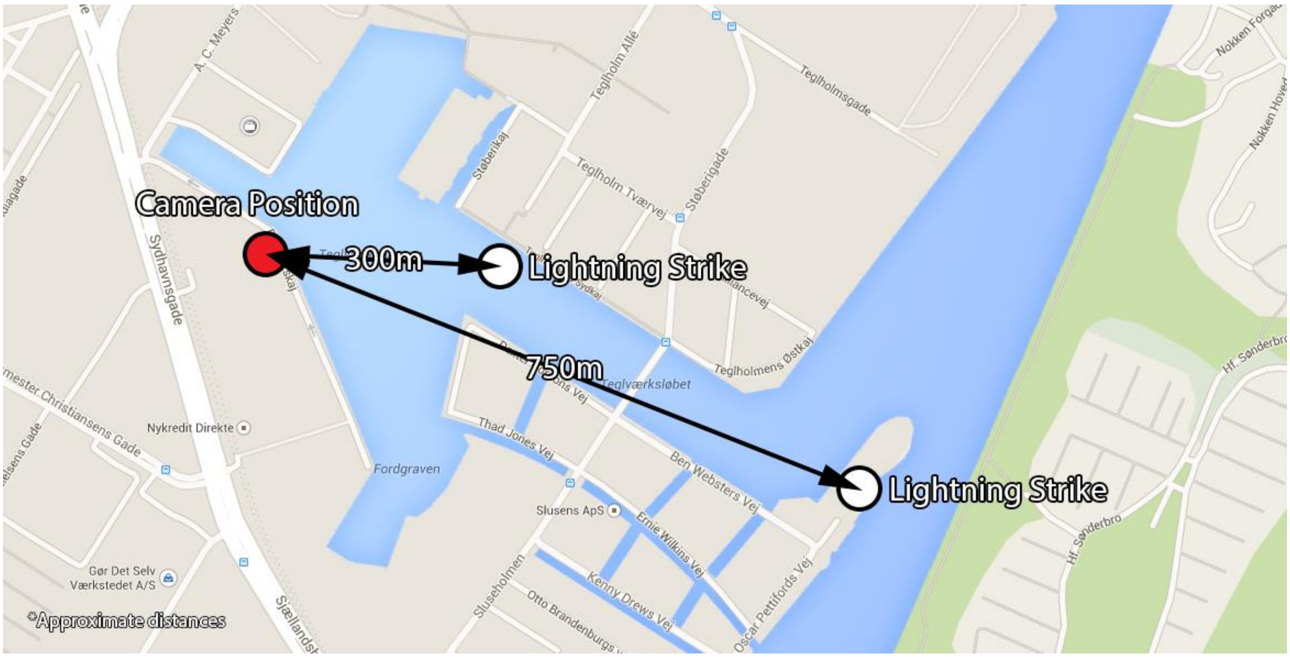


Figure 6 - Simulated locations of lightning strikes on real location

Below are still-images of each lightning strike in the two visuals created (Figure 7 & Figure 8)

Figure 7 - Simple Video, medium and far lightning strike

Figure 8- Rich Video, medium and far lightning strike

For the final implementation designed testing, this amplitude and delay was swapped between the two lightning strikes creating the incongruency needed for the test, alongside a control of the video with the sound and visuals matching.

The sequence was 40 seconds long, running at 25 frames per second in 1920x1080 resolution, with the audio sampled at 48 kHz.

Testing

Two test methods were used to obtain results for this experiment. The first being a personally conducted test, with participants being asked to answer questions orally, and their answer interpreted and compiled by the conductor. The second, an online questionnaire with the video embedded, asking the same question as the conductor in the first example.

In the conducted test, participants viewed one of the two versions of the experiment (with rich, or simple background), and upon the video ending they were immediately asked which lightning strike was closest. As the two strikes had their sound swapped (ie. Far strike with close sound and vice versa), their answer would describe which stimuli was dominant, as their binary choice would indicate if they determined the distance by visual or auditory perception. They were also asked to briefly elaborate on their choice

The video with the rich background had the first lightning strike as the closest visually, and the video with the simple background had second strike being closest to the camera. While the order was swapped, the strikes were identical, with the different backgrounds. Furthermore, the test featured a control video with matching audio visual stimuli, always being the other version (rich/simple) than the one tested with incongruent stimuli. The method and distribution of participants in the personally conducted test of experiment 1 is illustrated in Appendix A: Experiment 1 Conducted Interview Structure.

The online questionnaire featured the same video with rich or simple background, and participants were prompted with the same question after its completion, again making them pick which lightning strike was closest, and them briefly elaborating on their choice.

There were 16 participants in the personally conducted tests which were done at Aalborg University Copenhagen with the majority of participants being students. The online tests added 20 participants adding up to 36 participants in total.

Results

After correcting an erroneous answer from participant 10A from the online questionnaire, done on the basis of his elaborating comment, and subsequent answers, and finally discarding two participants from the personally conducted tests (reported incongruence as answer) there were a total of 34 answers.

The majority of participants (20 out of 34 - 59%) appeared to make their choice based on the tests auditory stimuli, reporting the closest strike as the one sounding closest, making this the dominant stimuli for these participants in this experiment (see Figure 9). Still, 14 participants appeared to make the choice from visual stimuli, meaning they chose the lightning strike that looked closest.

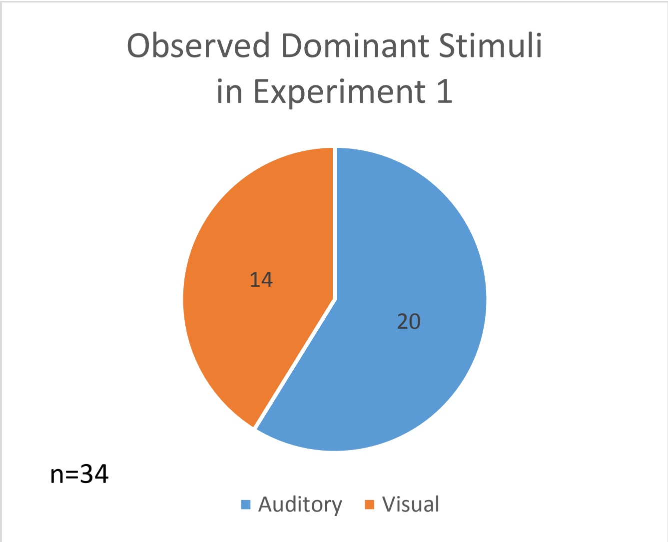


Figure 9 - Observed Dominant Stimuli in Experiment 1

The results of dominant stimuli did not appear to differ between the videos with rich and simple backgrounds when viewing the complete dataset. The video with rich background had an approximate distribution of 60% auditory dominance and 40% visual dominance interpreted, where the video with the simple background had a similar 56% auditory dominance and 44% visual dominance interpreted as seen in Figure 10.

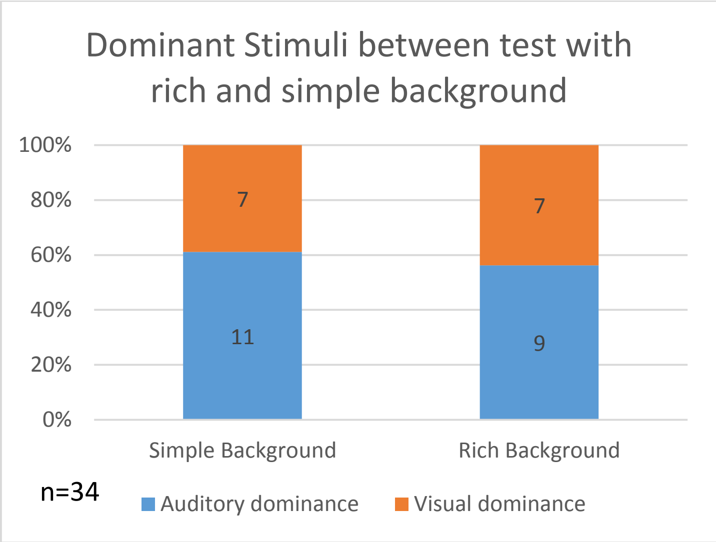


Figure 10 - Observed Dominant Stimuli between test with Rich and Simple Background

However, a difference was observed between the two testing methods (conducted and online). The interpreted dominance from the results obtained from the online questionnaire was 50/50%, where the interpreted dominance from the personally conducted test showed a larger majority making the choice based on auditory stimuli (71% auditory, 29% visual) (see Figure 11).

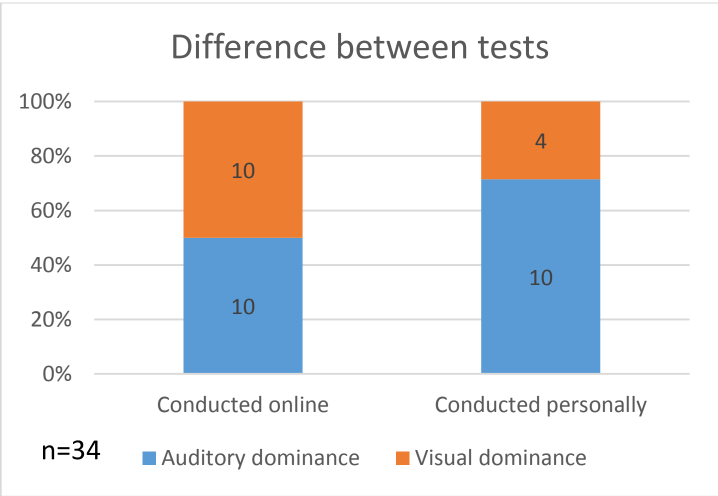


Figure 11 - Difference between tests in Experiment 1

The cause for this difference is as of writing unknown, but could be attributed a difference in monitor equipment from the conducted tests and those done by participants at their own equipment. While special care was taken to *not* reveal the question *before* the video in the online questionnaire, some participants may have actively analyzed the video potentially skewing the results, or they could have watched the video twice even when explained not to – but this is speculation at this point.

Experiment Conclusions

This experiment shows that distance to objects in video is determined by different stimuli between different people, with a slight majority having sound as the dominant stimuli (59%).

There were initially no indication that a rich or simple background had any effect on this perception of distance, but further analysis suggests issues in the results between the two test methods. Following participant comments, the rich background held more visual depth cues that the simple background did not have, such as occlusion and texture gradient, potentially making visual stimuli more effective in rich scenes, and auditory more important in simple scenes.

The null-hypothesis is accepted as the experiment did not show a significant difference between simple and rich visual backgrounds in video with a p-value of 0,855 ($p > 0,05$), .

> **H_0:** There is no significant difference in perception of distance in audiovisual with rich or simple backgrounds.

The experiment can conclude that participants perceived distance with different stimuli, and making both important for the creation of synthetic elements in video. While there was no clear dominant stimuli, participants did appear to favor determining distance based on auditory cues.

Experiment 2: Manipulating Perception of Direction and Trajectory

As a continuation to Experiment one, additional interest was raised in localization of a time moving sound source. As previously stated (Perceiving distance with sound) it is easier to determine the location of a sound source that is moving, compared to one that is not, which is also true for moving visual elements. Again, congruency was of focus, investigating what happens when visual and auditory stimulation display incongruent stimuli.

Combining incongruent stimuli can lead to changed perception and even audiovisual illusions, which is described in the theory of Synchresis.

Synchresis

The theory of Synchresis proposed by Michel Chion explains the interaction in the connection between film-sound and video. Specific auditorial stimuli can alter our visual perception, effectively creating illusions making us think we see something that isn't there, because of the sound alone (Chion, 1994). The function of film sound is to support and affirm the video and moving pictures, further animating them. This can be called a synchronization of the senses where one intensifies the other, creating deeper understanding, narrative and ultimately a unity between scenes, story, shots, and video in its entirety (Carlsson, 1994) .

This experiment wanted to investigate movement and sound as these are perceptually tied very closely together, because the perception of sound itself signifies movement or agitation of one kind or another. When we perceive something moving visually and determine it as a source of a sound, this information is not worked on in the brain, instead the visual observation and auditory perception is regarded as a single event. This unification is then Synchresis being the fusion of "synchronization" and "Synthesis" (Chion, 1994).

As this sound and movement is perceived as a single fused event, *"the sensory impressions are overlaid so that picture and sound contribute to the nature of each other"* (Carlsson, 1994). This means that a video of a person closing a door normally overlaid with the sound of a door being slammed creates the visual illusion that the door is being slammed, even though it isn't visually. This indicates that sound is often a dominant stimuli, and it is this hypothesis that the upcoming experiment is investigating.

Changing the perception with perceptual synchresis is a known and often used method in film and video, where the sound editor can change all manners of perceptions of visual objects – make fake props seem heavy and real, adding "swooshes" and impacts to punches making them seem more powerful and the like.

Experiment 2 Overview

The question was if auditori stimuli was sufficiently dominant to change or affect the visual perception of the position of an object and its trajectory. For this, the experiment focused on an object flying towards the camera appearing on a trajectory that would hit it, while the sound will be appear not to. The video was then stopped at some distance and time before the object hits the camera to maintain some level of ambiguity, and the participants asked if they thought the object would or wouldn't hit the camera. This was tested with an offset sound and a congruent sound to the object for testing and control.

Practically, this experiment needed an object that continuously emits a sound which was believably able to travel through the air at some speed and control. For this, an airplane was first proposed, which then evolved into a science fiction space craft, as the object was required to have very high maneuverability and a very distinguishable sound. The well-known "Tie-Fighter" from Star Wars was chosen for its supposed familiarity with participants, and because of its signature sound as created by Ben Burrt.

This experiment also falls within the category of intersensory bias and adaptation (Biocca, et al., 2001) for its manipulation and changing of incongruent stimuli, but it also touches upon the category of Cross-modal enhancement or modification. This was due to its changing and enhancing the visual stimuli with the auditory, especially in a control version where the sound was to amplify the trajectory of craft, potentially hiding some lower quality visual elements with high fidelity audio.

The video created for this experiment contains:

- Static camera filming a harbor front.
- TIE Fighter rounding a corner some distance away and flying towards the camera
- Picture going black when the fighter was approximately 30m away.
- Sound of fighter constructed in surround sound, and intentionally moved to the side in one version of the experiment.

More specifically, the video shows a bridge between two buildings (at AAU-CPH's campus) that the Tie fighter flies underneath, situated at the same harbor area as Experiment 1 (see Figure 12). The video was created with two different audio tracks; one with congruent sound matching the trajectory of the craft, and one with the sound drastically panned to one side in the end. The goal of the experiment was to investigate if incongruent auditory and visual stimuli in video can still appear as a unified perception and posed following hypothesis:

H_1: Incongruent auditory stimuli have a significant effect on the perception of trajectory of an audiovisual object in video.

Design & Implementation

As mentioned, a Tie-Fighter from the Star Wars films were chosen as the moving object because of its familiarity and signature sound. This fictitious craft was 6.4m long and capable of travelling 1200km/t (wookiepedia, 2014). The size was simulated, while the speed was lowered as the craft makes a sharp turn.

The sound was that of the original movies composed by Ben Burtt, a now well-renowned sound designer that is regarded as one of pioneers within film sound and effects. This TIE fighter's sound is composited of the unlikely elements of an elephant's call and a car skidding out on wet asphalt (Rinzler, 2001). These sources come together to create a very open, loud and menacing sound that was very different from the harbor's other sounds and soundscape.

To further "sell" the illusion of a moving craft over the water, additional effects was necessary for a successful integration of the rendered and real footage. This includes a reflection in the water, as the craft passed closely over it, as well as water spray from the surface as it passed over it.

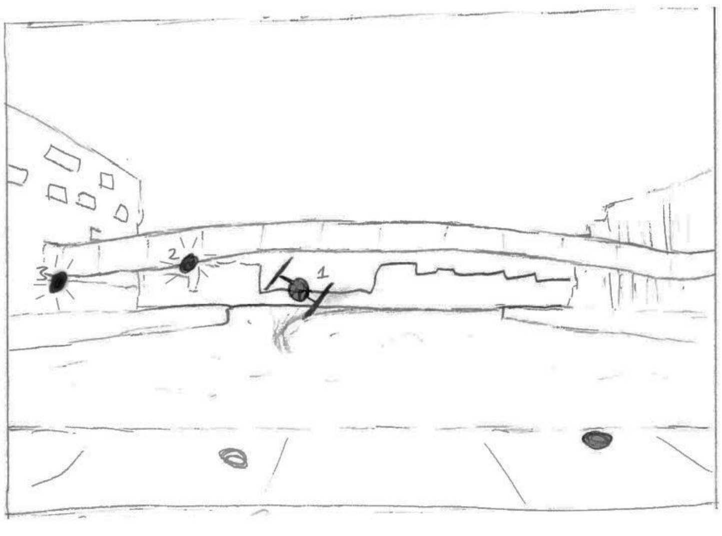


Figure 12 - Initial concept of Experiment 2 - 1. Object; 2/3. Sound Source

The model of the tie fighter was obtained and later modified from a database known as SciFi3d that holds a large collection of publicly created and available models from various sci-fi universes. Credit goes to James Bassett for the creation of the original model which was modified for this experiment with a change in textures for faster rendering and general style (Bassett, 2000).

The model was modified and animated in "3Dstudio MAX" In which reflections and water spray was also created. The craft was made to scale and animated to follow a path that would lead it around a building to the right and then straight towards the camera (see yellow line, top-right of Figure 13).

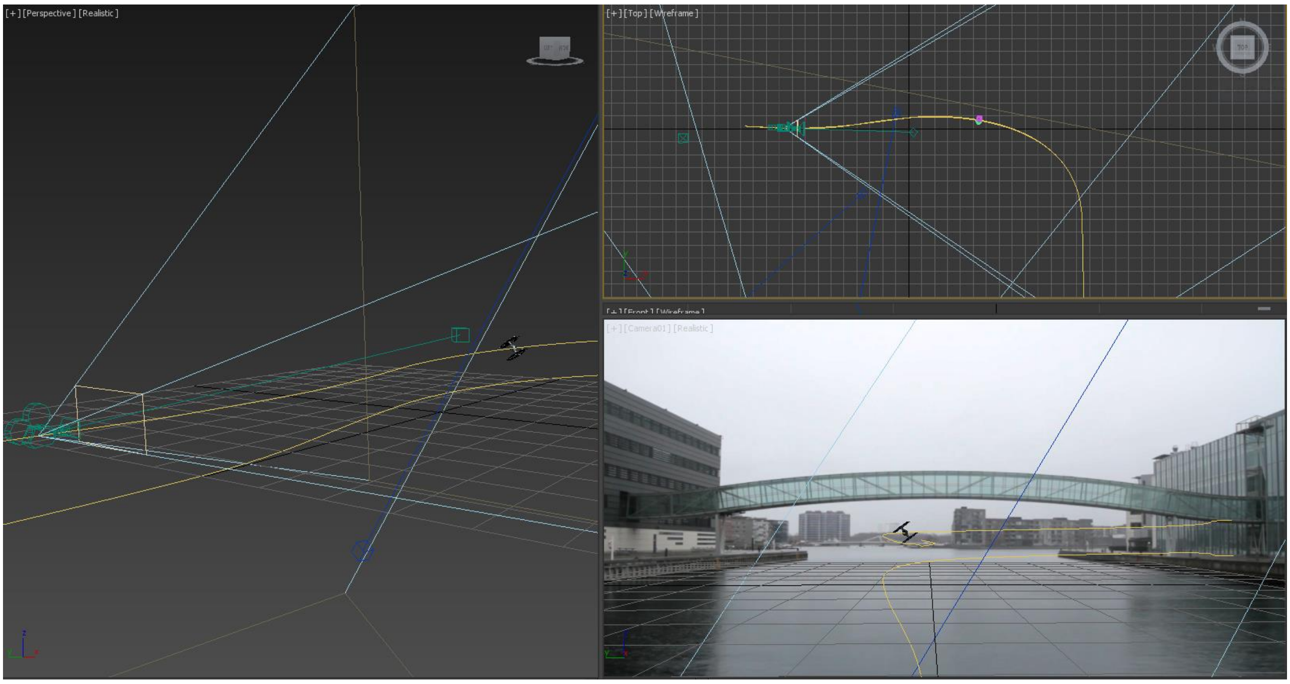


Figure 13 - Animation of Craft. Yellow line being the animation path of the craft and water spray.

Following the trajectory, the craft banks and rotates along the path as if it was an aircraft, as the craft do in the movies, firing two laser beams from its weapons toward the end. The clip ends 1-2 seconds before the craft would hit the camera and just after it shooting at it. As visible above (Figure 13) the craft moves in an almost straight line towards the end, clearly indicating visually that it will hit the camera, while still having some movement in the animation helping a more natural behavior of the craft as it straightens up and corrects its path.

The model was rendered with motion-blur and the various visual elements (model, reflection, spray) were rendered separately and composited thereafter. The elements were graded individually, to integrate with the real footage.

As mentioned, the sound for the craft was made in two versions resulting in two variants of the scene. The sound was made following the animation, and created in 5.1 surround with the source of the sound following the approximate path of the craft (See Figure 14). An alternative version of the sound was made where the end of the sound pulled substantially to the right, which was coupled with the same straight-on visuals to test which was dominant (visuals or sound).

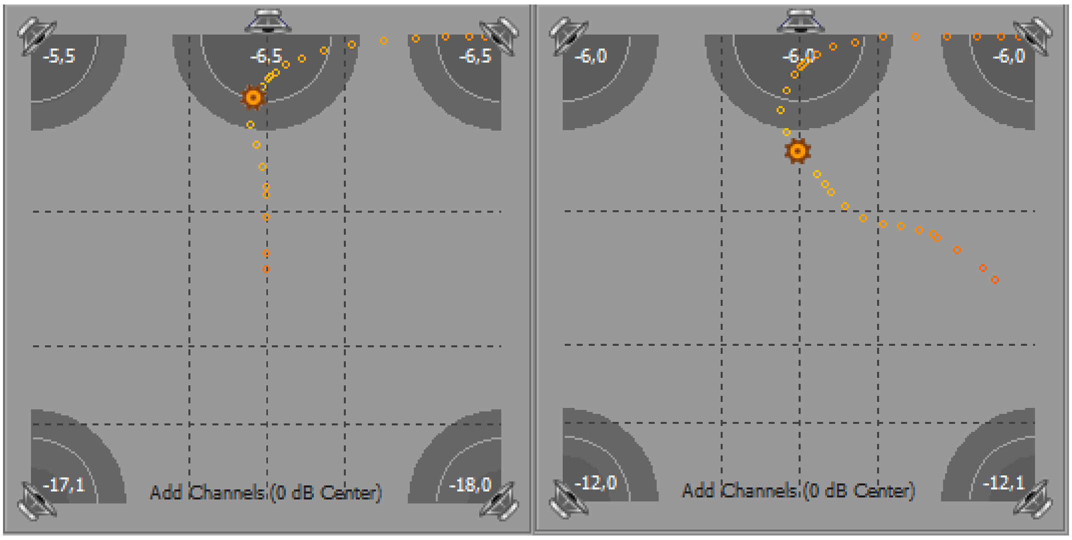


Figure 14 – Surround pan of the track with the craft sound – congruent (left) and incongruent (right) variant

The sound itself was made from three variants. One sounding far away, one sounding a more stressed, and one up close, composited in that order fading from one to the next. Additionally, right before the end of the shot where the craft shoots two laser beams, these are also coupled with the craft's signature laser sound.

Furthermore the amplitude was keyed such that the craft gets progressively louder as it comes closer, with an increase as it rounds the corner of the building, with a clear line of sight and sound. An open, but audible reverb was also added simulating the sound being reflected off the large buildings.

Figure 15 - Experiment 2 still image, final look

The sequence was 28 seconds long, running at 25 frames per second in 1920x1080 resolution, with the audio sampled at 48 kHz.

Testing

As with the previous experiment, testing was done with two methods; personally conducted and online questionnaire. 16 participants were included in from the personally conducted test, and 20 participants from the online test.

In both cases, participants were asked to watch the video of the craft moving towards the camera, and when the video was stopped prematurely, they were asked if the craft would hit the camera if this video had continued (either physically or with its weapons that fired right before the video stopped). Afterwards they were asked to briefly elaborate on their answer, for additional qualitative data.

When participants gave their answer whether it would hit the camera or pass the camera, this would give an indication to if they made the decision based on auditory or visual stimuli, along with their elaborative comments to the choice.

No participants or results were discarded from this test.

Results

The 36 pair of answers from the participants resulted in a 14/12 distribution of results, where 12 stated the craft would hit, and 14 stated that it wouldn't. As such there appears to be no universal dominant stimuli in determining the trajectory of an audiovisual object in video, as instead different people may rely on different stimuli for this perception of trajectory and direction (see Figure 16).

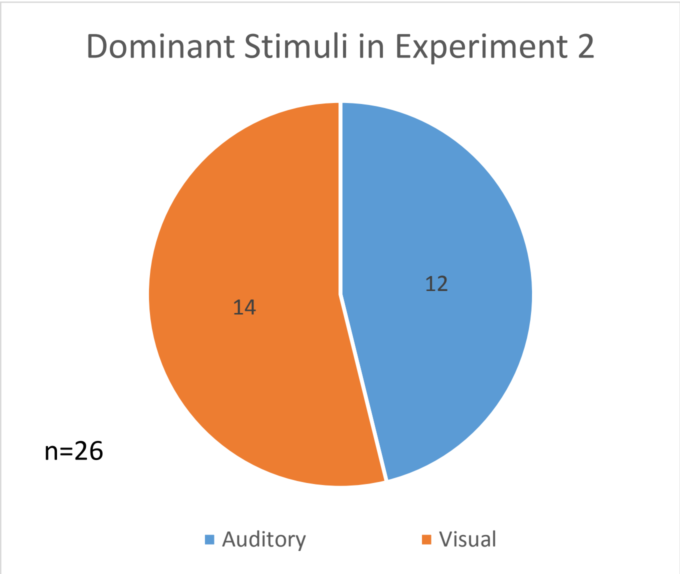


Figure 16 - Observed Dominant Stimuli in Experiment 2

Some elaborations from participants described how they said the craft or laser went to the right, despite the graphics being centered. *"It looked like it was to the right…"* (P11) and *"The laser went to the right"* (P3). This altered perception is from auditory stimuli which was moved to the right, proving that auditory stimuli can dominate and even change visual stimulation.

When looking at the test with the audio shifted in the end and the control with the audio matching, there is an observed difference (see Figure 17). The majority of participants in the control said that the craft would hit them (20/26) as expected with the two stimuli being congruent. But as mentioned above, the test showed 12 participants reporting it hit, and 14 that reporting missed, having more responses saying it missed than the control.

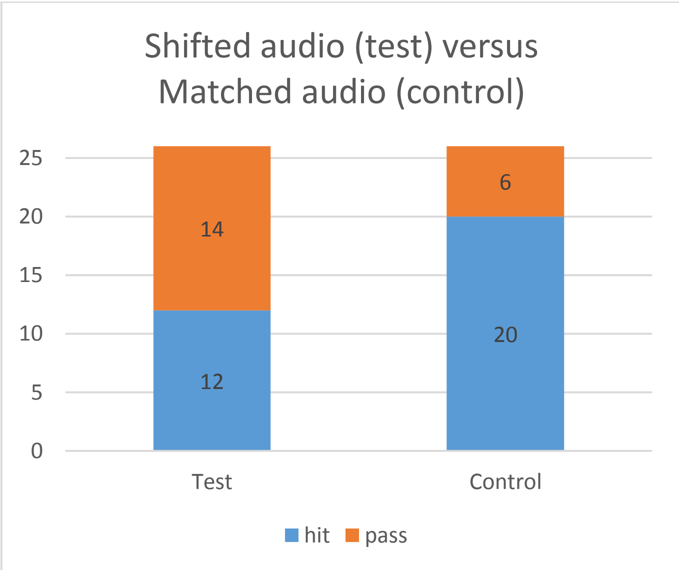


Figure 17 - Test (shifted audio) vs. Control (matched audio)

As was also the case with Experiment 1, there was a difference between the personally conducted test and the online test, this time with a greater disparity between the two (see Figure 18). In the results from the online test, participants appear to make the choice based on visual stimuli (13 of 20), where the conducted test show the majority of participants relying on auditory stimuli (11 of 16).

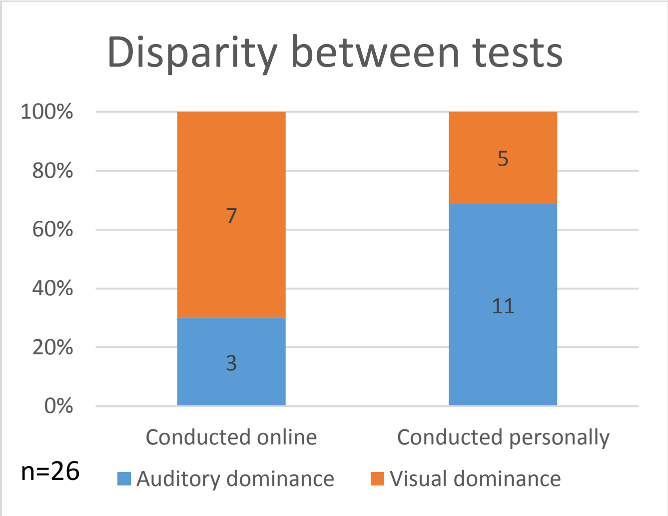


Figure 18 - Disparity between test methods in Experiment 2

The cause for this disparity is as of writing still unknown, but again different monitor equipment between tests is suspected to have influence. Likewise, participants may have intentionally analyzed the video or watched it twice in the online questionnaire which could also change the results.

Experiment Conclusion

This experiment illustrated the theory of synchresis as the auditory stimuli made some participants perceive visual events that were not featured in the video (mostly from answers from the conducted test). Interesting to note was the fact that the visuals had little-to-no ambiguity with their regard to trajectory, and that this clear visual stimuli of the craft hitting the camera was sometimes overruled by the auditory stimuli of the craft moving to the right.

As with experiment 1, auditory stimuli proved capable of determining and overriding the perceived location of objects, this time with a moving source – although it did not prove to be better or more dominant than visual stimuli.

The null-hypothesis was rejected as results showed a significant difference between the test and the control with a p-value of 0,022 ($p < 0,05$), suggesting that incongruent auditory stimuli had an effect on our perception of trajectory of the audiovisual object.

> **H_0:** Incongruent auditory stimuli does not have a significant effect on the perception of a trajectory of an audiovisual object in video.

This experiment indicated that sound efficiently affects our perception of an object and its trajectory in video, making sound design an important aspect when creating moving CG audiovisual elements for video.

Experiment 3: the Motion Bounce Illusion in Rich Media

For this study and experiment, a range of audio-visual illusions were looked upon for inspiration and implementation into video. Illusions were regarded as an interesting and effective way to display and test audio visual interactions and the interest was mainly with crossmodal theories and stimulation.

A very popular crossmodal audio-visual theory is the previously mentioned 'motion bounce illusion' in which two similar 2D graphical objects can be seen as intersecting or bouncing off each other, the perception being influenced (towards bouncing) if a sound is played at the point of intersection. The popularity of the illusion is likely because it demonstrates a simple and effective interaction between audio and visual stimuli, making it interesting for this investigation.

The visual stimuli of the illusion has no bias towards the visual objects either bouncing or intersecting, suggesting any testing with participants would result in random perception of them passing through or bouncing. This was however not the case, as the majority participants report the objects as bouncing if a sound was played at the intersection, and likewise the majority reports them passing through each other without the sound (Sekuler R, 1997). With the scene always remaining entirely ambiguous, these results are surprising, but for this study the question was if this illusion only occurs in experiments with simple visuals and sound. Previous studies with the illusion have been with simple flat shapes with straight trajectories, and the sounds used at the intersection has been simple sinusoidal beeps and tones (Sanabria, et al., 2004).

What happens if the 2D visual elements are replaces with real 3D objects, if the sound at the intersection is congruent with the 3D object's real sound and if all of this occurs in real space filmed on video, including realistic trajectories for the objects as they are affected by gravity? It was clear that a lot of care has to be put into the production of video that both demonstrates this illusion while maintaining a photorealistic look.

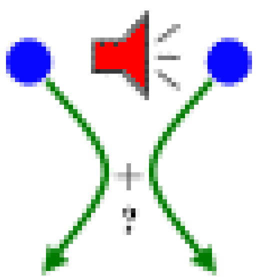


Figure 19 - Motion Bounce illusion as illustrated by Michael Bach

Trajectory

If the colliding or intersecting objects were to be animated for photorealistic video, their animated trajectories had to follow realistic paths affected by gravity, wind resistance and initial applied force.

An object thrown in vacuum with gravity will follow a parable (see Figure 20, blue line, top figure), but with air resistance or drag this trajectory changes with the object losing some of its horizontal (and vertical) velocity over distance (see Figure 20) causing the arc to change significantly. (Gibbs, 2013)

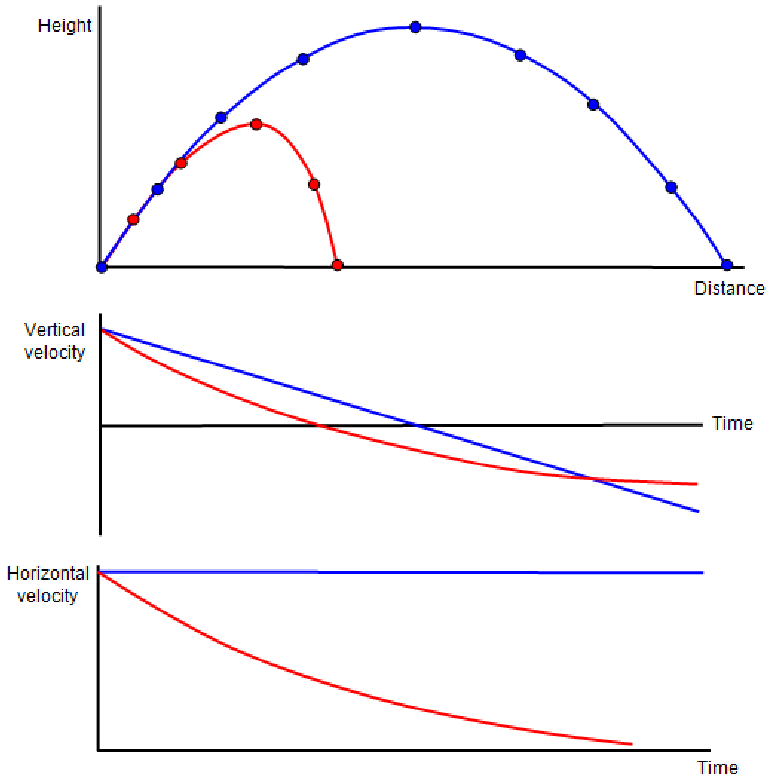


Figure 20 – Trajectory and velocity of object with air resistance (red) and without air resistance (blue) - (Gibbs, 2013).

As this experiment was to have objects moving in a realistic fashion, their path and trajectory in air had to follow these physical principles for a believable look.

For this experiment, and application to video in general, it was important to note that physical aspects such as these mentioned is not required to achieve realism. It has to look realistic, not *be* realistic – which is a big difference. Video and film often strikes a fine balance between the possible and impossible with cars flying through the air and the like for dramatic effect. Events are made to look plausible, not completely realistic and in this process, many laws of physics are omitted or not necessary and can even be detrimental to a scene by removing control from the director or animator, which gives this experiment freedom to manipulate these trajectories, as long as the look realistic.

Experiment 3 Overview

Converting the 2D illusion to 3D took special care and attention to detail. First of which, was to approximate the 3D environment to 2D in such a way the illusion would work. Depth of the image should be hard to discern for the animated objects, as should occlusion meaning that it shouldn't be visibly possible to see which object was in front of which.

Depth was made ambiguous by having objects intersect in a section of sky, with as few depth cues as possible, and it was possible to hide occlusion in motion-blur with the objects in one version. Likewise, it was presumed favorable to have the intersection with either small objects, or at some distance to further avoid smaller details of depth or occlusion. The objects intersecting produced a familiar and clear sound and was made to be audible over any background-noise of the area. As with previous experiments, this aimed to have some meaningful content as video including the test element. Following the thesis' pursuit of high production quality and its desire to implement theories to practical implementations, this experiment's test and product featured a small narrative.

To fulfil the experiment's criteria the following scene was established:

- Video designed as an improvised recording of a couple of guys trying to achieve something very difficult
- Camera initially handheld, later lying on the ground filming upwards (30 degrees).
- Two people claiming to have "a superior aim" are featured.
- They proceed to walk quite some distance away from each other, each with a tennis-ball.
- They throw the tennis balls towards a point in the sky, where they collide or pass twice.

The tennis balls were animated manually after leaving the hand of the two persons to precisely pass where intended. A sound of tennis balls was present or not present when they intersected.

The video was made so that the two persons throw the balls two times – with ambiguous reactions between each throw. The sound of the tennis balls colliding was present in only one throw.

This video and test was created with and without motion blur thus raising the following hypotheses:

> **H_1:** The motion bounce illusion can successfully be implemented in realistic video
> **H_1:** The addition of motion blur makes a significant difference on the likelihood that the motion bounce illusion is percieved

Design & Implementation

As mentioned, a scene was shot at an open area with two actors, each with a tennis ball. The clip was made as an improvised recording of a couple of guys wanting to capture a cool trick on video. With the difficulty and unlikeliness of two tennis balls hitting in mid-air with some distance between the two people, the point was that the balls hit at their first attempt. The camera was handheld (poorly shot in general on purpose) as they move out into the open space, where after the camera was placed on the ground shooting in profile, having both actors in frame. The scene was shot in one take to simulate unedited footage, deliberately with somewhat poor quality camera work.

As mentioned, the original tennis balls thrown were removed after they left the actor's hand, using the background from the previous frame painted over.

Surprisingly, when analyzing the recorded shots for implementation, the real tennis-balls had no motion-blur (see Figure 21) as they travelled between the actors. This was a minor oversight, as the shot was of course filmed against a very bright sky, requiring a very fast shutter speed on the camera, effectively eliminating any motion-blur that was originally expected.

Figure 21 – Still frame from raw footage. Tennis ball in-air moving right, having no motion blur

The animated tennis ball was made from a picture taken same day and place (see Appendix B: Picture of Tennis ball used for Experiment 3), and then animated to follow a plausible intersecting trajectory, created on the basis the previous theory of an object being thrown with wind resistance. This animation had several requirements such as arrival at the other person at the right time as he catches it, moving slowest at the top of the arch, and intersecting the same place and time to facilitating the motion bounce illusion (see Figure 22).

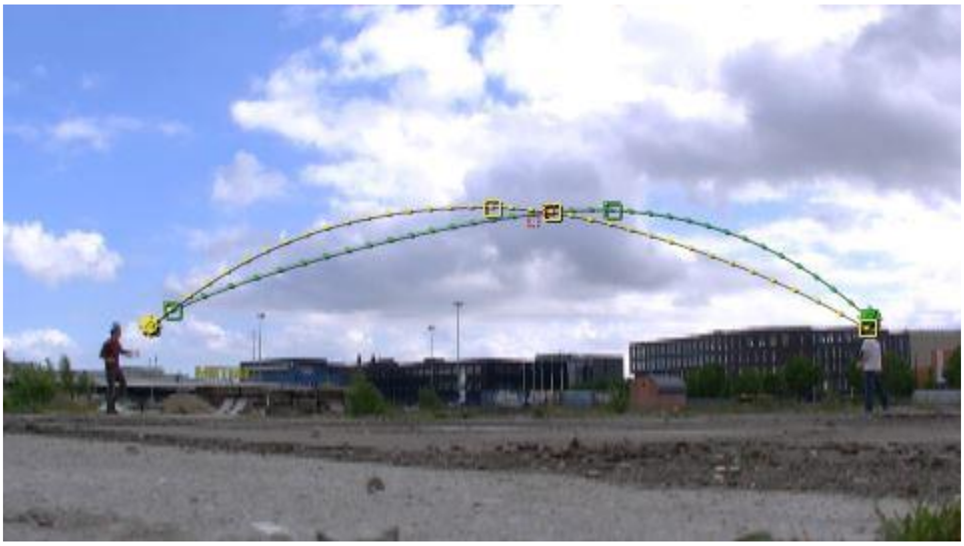


Figure 22 - Trajectories of the two tennis balls, first throw - Yellow thrown from right, green thrown from left.

Two versions of the video was made for experimentation; with and without motion blur. This was to investigate if motion blur has any effect on the motion bounce illusion when created with photorealistic video. The video clips was identical apart from the motion blur, with two attempts in each – one with, and one without a sound of the tennis-balls hitting.

The sound of the tennis balls colliding was recorded in an anechoic chamber with a RØDE NTG-3 microphone connected to a zoom-recorder. The tennis balls were thrown together and the sound of the collision isolated and treated with a band-pass filter for the video.

The sequence was 54 seconds long, running at 25 frames per second in 1920x1080 resolution.

Testing

Participants were shown the video in its entirety with its little narrative and immediately asked (when it was done) how many times the tennis balls hit each other in mid-air. They then had to select "they hit the first attempt", "they hit the second attempt", both, or none. As mentioned, a sound was present at the first intersection, and if participants chose the option of a hit in the first attempt, the motion-bounce illusion was successful.

Testing was carried out with an online questionnaire with the video embedded. Participants would watch a version with or without motion-blur on the tennis ball.

24 answers were collected, from which none was discarded.

Results

Of the 24 answers, 22 participants reported that the tennis balls only hit the first attempt, and 2 participants reported it only hit the second attempt. There were no reports of them never hitting or hitting at both attempts. This proves that the motion bounce illusion was achieved successfully, and that it is possible to create the motion bounce illusion with photorealistic visuals and real sound recordings.

There was no significant difference between the video with motion blur and the one without (p-value of 0,167 > 0,05), but there were two reports of the tennis balls hitting the second attempt in the experiment with no motion blur, being the attempt with no sound at the intersection (see Figure 23). As mentioned this is not enough to prove that motion blur as any effect on the motion bounce illusion in video.

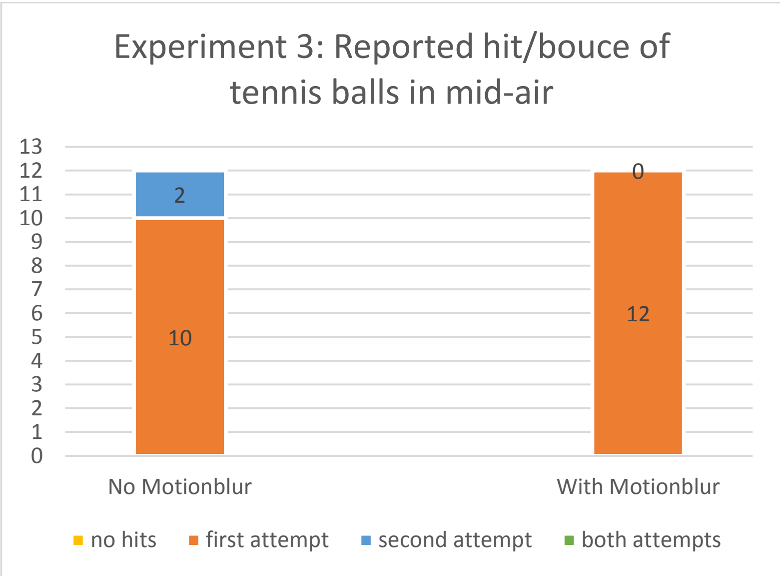


Figure 23 - Reported hit/bounce of tennis-balls in mid-air in Experiment 3

Experiment Conclusions:

This experiment can conclude that the motion bounce illusion is possible to implement in realistic video, and therefore the first null-hypothesis is rejected:

> **H_0:** The motion bounce illusion cannot be successfully implemented in realistic video

However, the second null-hypothesis is accepted as there was no significant difference between the test with and without motion blur, with a p-value of 0,166 ($p > 0,05$).

> **H_0:** The addition of motion blur makes no significant difference on the likelihood that the motion bounce illusion is perceived.

Experiment 4: Expectation following Obscured Audio-Visual Event

This experiment originates from a couple of brief conversations that happened before a preliminary testing of the first two experiments. With the previous experiments in one way or another dealing with a congruency of audiovisual elements, they did not touch upon the element of 'expectation', which also ties very closely to congruency.

The interest in expectation started during the testing of the second experiment (Experiment 2: Manipulating Perception of Direction), where a foreign/alien sound was introduced in a realistic soundscape. This sound was of the moving object in the experiment (the TIE-fighter) having distinct and aggressive auditory characteristics, all taking place in a quiet harbor area at the university. Such incongruent auditory information was unexpected, and because of this it generated a lot of attention, and potentially even alarm or alert (Carlesa, et al., 1999).

Having incongruent or alien auditory stimuli is unexpected in terms of our perception, which may lead to different interpretations and perception of a scene of audio-visual elements. "*Change in sound-image compatibility conditions is enough to produce quite different aesthetic and affective reactions."* (Carlesa, et al., 1999).

Our expectations while viewing video can, and has been effectively utilized to create surprising, unsettling or humorous moments in film and video, just to name some of the applications. As an extra experiment for this study, interest was raised in this expectation and its connections with sound, and a simple questions was posed for investigation. Does immediate auditory stimuli have a significant effect on the expected outcome of an obscured audio-visual event?

This experiment was to be considered as an additional and extra investigation, as it primarily serves to confirm and validate a personal interest in the area of expectation and audio. As such, this experiment was of smaller size and significance, but still relevant to the study as a whole because of its inherent audio-visual interaction.

Experiment 4 Overview

To test this expectation following an obscured audiovisual event, a little scene was created taking place at a nondescript indoor location around a corner. Again, a small narrative was created to fulfill the investigations pursuit of practical application onto video of high production quality.

These are the key points of the experiment video:

- A person runs around a corner with the camera following throughout the shot
- A box was placed upon a locked door (with tape)
- The person goes back and braces for an explosion
- This explosion occurs with debris and sparks flying past him

The sound of this explosion was created in two versions. One loud and exaggerated, and one more quiet and sounding smaller. This was to examine of the sound of the explosion changes our expectation of damage inflicted by it. As with previous experiments, the visuals of the explosion remain the same.

This experiment poses the following hypothesis:

> **H_1:** Changing the sound effect of an obscured audiovisual event makes a significant difference on the viewer's expectation of the result of the event.

Design & Implementation

As in experiment 3, the camera was handheld, and the scene was done in one shot. Where the previous experiment had the camera stationary during the manipulation of the picture, and where all other previous experiments in the thesis had static cameras throughout, this experiment had the camera handheld during the frames for compositing, during the explosion. This meant that the motion of the camera had to be tracked and mapped to every composited element. A process in which a small mistake could compromise the element of realism and break the illusion of the explosion.

While it was possible to track and emulate the camera, this process is time-consuming and prone to errors. Instead a select few elements of the image around the location of the explosion were tracked and their X-Y position outputted to carrier object (a null object in the software Adobe After Effects). The elements of the explosion were then parented this object, inheriting its movement in the X-Y plane. This technique was only viable as the camera remained in roughly the same position and angle throughout the shot.

As is sometimes the case with compositing (from personal experience in professional compositing), the thorough and realistic way of adding effects to a scene is not always the best or most successful. The task for this experiment was to add an explosion to a scene. The realistic 1:1 way of creating this effect would be to simulate all aspects in fluid, smoke and particle simulations in 3D software, creating smoke, fragments and fire as the effects. Matching reality with such an approach is a difficult and complicated process, as well as time consuming. Instead, compositors often choose to use real footage shot on black background (converted to a transparent alpha channel), as did this experiment. For this experiment, 3 shots of real elements were used, and one of simulated particles. The fact that these shots were of real events means that they also look real, a difficult look to get exactly right with simulation of computer generated elements.

These elements were a small explosive charge viewed in profile, a layer of dust being affected by an explosion and one of viscous fluid dropped in water – the latter used as smoke traveling along the floor of the scene. To have fragments of the wall/door affected by the explosion, some debris was present in the dust element, which were supplemented with a few simulated and rendered elements to have enough to make the wall seem sufficiently affected (see middle image of Figure 24).

Figure 24 - Compositing of Experiment 4. Foreground and explosion elements, frame 366

In order to have this explosion event happen behind the actor in the scene, he and the wall of the corner had to be separated from the background and placed in front of the explosion. For this, the moving actor had to be masked out in the moving handheld shot from the camera. This was achieved with a mix of manually tracing the outline of the actor, and using some automated processes similar to tools used in image manipulation for purposes of selection. Following the elements were placed in layers (see Figure 24) resulting in the finished result as seen in Figure 25.

Note that the elements of the explosion (middle image of Figure 24) does not correspond with where the explosive charge was placed, nor do they match with each other when standing alone. However, in the finished composited image, the elements unify well as a single event. This was intentional and made sure all elements were presented as desired. Not present in the still image below (Figure 24) is an illumination of the wall behind the actor.

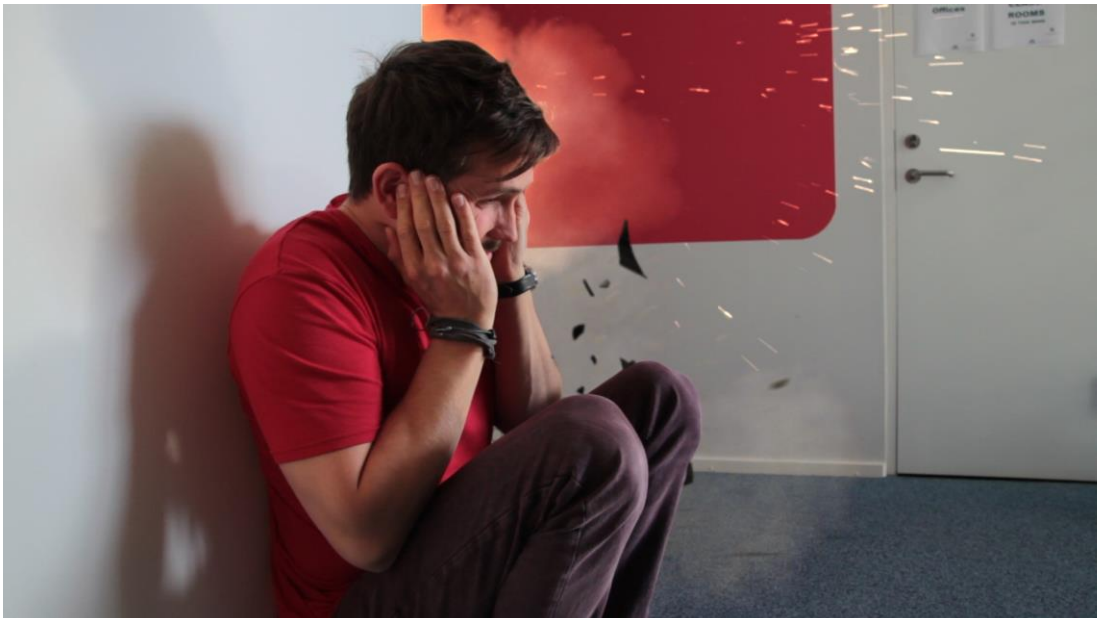


Figure 25 - Finished compositing result of Experiment 4, frame 366.

The sequence was 18 seconds long, running at 25 frames per second in 1920x1080 resolution. A standard within video production.

The two versions of the sound in the experiment were as mentioned created to portray different sizes of the explosion. The one used for the large explosion consisted of audio recordings of explosions, and a building collapsing. This was supplemented with the sound of debris landing and rolling together with concrete and wood breaking. With this the final sound had a sharp attack, very high amplitude over a long period with slow decay (see Figure 26).

For the smaller sounding explosion, a single strike from a bullet into concrete was used together with a lighter sound of debris and wood breaking and rolling. A short composited element of a brief impact created for use in motion graphic was also used to give the sound more depth. The result was a sound with sharp attack, and faster smoother decay, with a lower amplitude (see Figure 26).

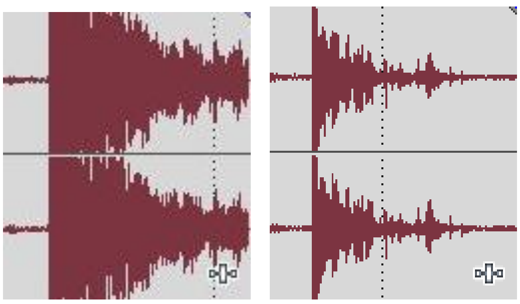


Figure 26 - Final sound composite of Experiment 4 in waveform. (left: large explosion, right: small explosion)

The timbre between the two sounds are their main difference, as one sounds like a small building being blown up and coming down, while the other sounds like large firecracker breaking something.

Testing

As mentioned, two versions of the explosion sound was made, one for each variant of the video. Participants would then watch one or the other variant in an online questionnaire. When it was done they were asked to guess the damage done to the door by the obscured the explosion. 6 options were provided from which the participant had to pick one:

- The door is unaffected
- The door has taken some damage, but there is no penetration (no hole)
- There is a small hole in the door, perhaps an arm could fit through
- There is a large hole in the door, one might be able to crawl through (awkwardly)
- The door is gone, only a bit of wall left around it
- The entire wall is gone

25 participants were included in the test, with 13 viewing the video with the small explosion sound and 12 viewing the video with the big explosion sound.

Results

The majority of selected answers to the small sound explosion indicated an expectation of a "small hole" or less damage as seen in Figure 27.

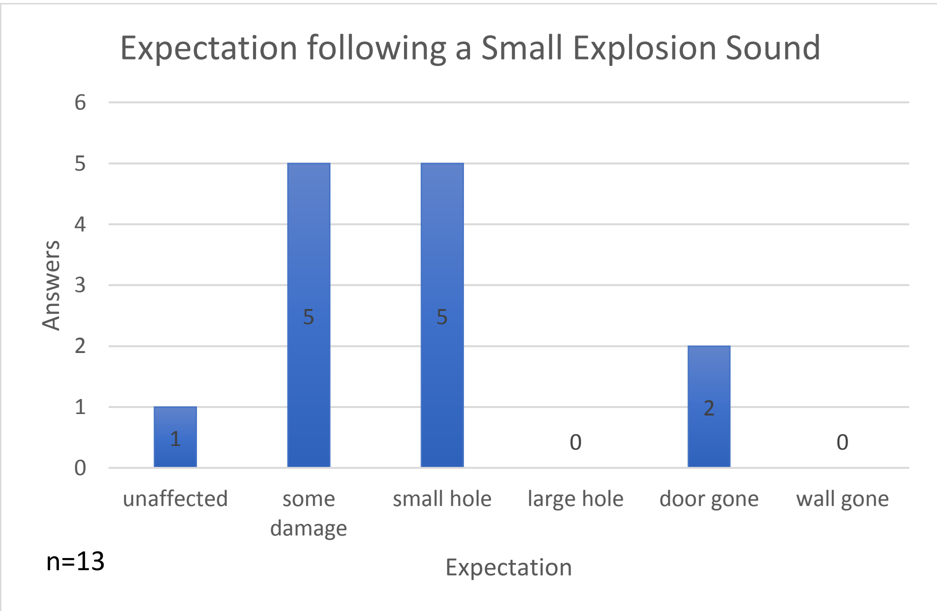


Figure 27 – Reported expectation following the video with a small explosion sound

The majority of selected answers to the large sound explosion indicated an expectation of a "small hole" or more damage as seen in Figure 28.

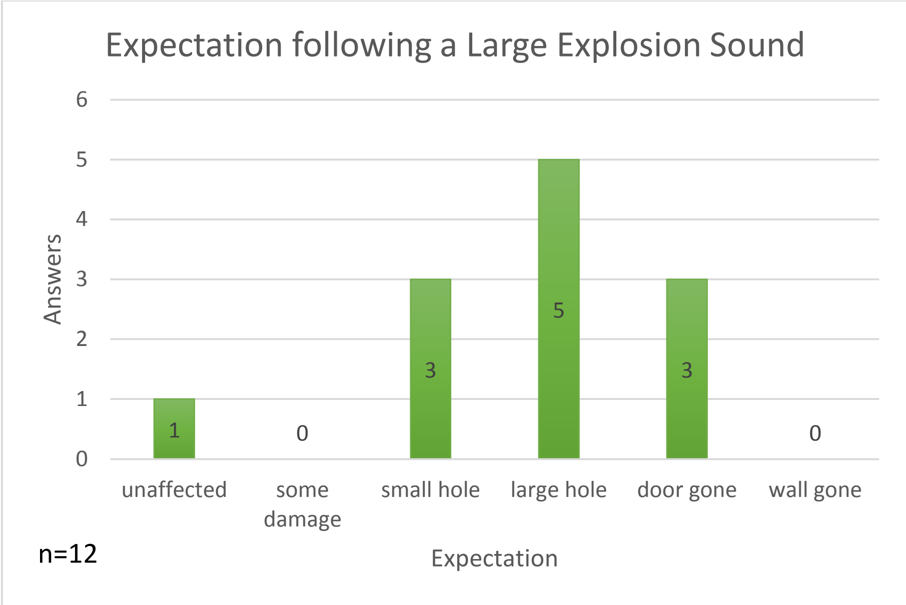


Figure 28 – Reported expectation following the video with a large explosion sound

Experiment Conclusions

With the results above one can conclude that there is a significant difference in expectation when the sound effect is changed to partly obscured audiovisual event. This is backed by a p-value of 0,044 ($p<0.05$) proving the significant difference and rejecting the null-hypothesis:

> **H_0:** Changing the sound effect of an obscured audiovisual event makes no significant difference on the viewer's expectation of the result of this event.

Experiment 5: Combining knowledge and elements from previous experiments.

The fifth experiment was an attempt to combine elements from previous tests into a single product of high quality and rich details. This was not meant as a larger final product, but rather a summary of knowledge learned from the previous experiments and implementation.

Experiment 5 was designed to incorporate at least one main element from each previous experiment and their conclusions, and test these in unity with realism and perceived quality reported by participants. Due to the previous tests dealing with events occurring in highly specific examples, this test was designed with a lot of synthesized elements to facilitate these extraordinary circumstances.

While the experiment wasn't designed as a final product, it *was* meant as a "finale" of the thesis, and received appropriate attention in implementation because of it.

Experiment Overview:

Again, a small narrative was chosen as led by a handheld camera observing elements derived from previous experiments, combining them into a single scene and extended event.

What follows is an overview of elements from each experiment implemented into experiment 5:

- Experiment 1
    - Sound and distance - Auditory depth cues
    - Dominance and importance of sound.

- Experiment 2
    - Sound and trajectory
    - Spacecraft as solution to a difficult implementation

- Experiment 3
    - Motion bounce being possible in rich media.
    - Improvised video look had positive response

- Experiment 4
    - Expectation of the explosion.
    - Camera tracking basis

As mentioned the camera was handheld as in experiment 3 and 4, continuing a successful improvised look to the video that started in experiment 3. In this fifth experiment, the camera and its holder emerged as running onto a rooftop where the camera was immediately pointed at the sky. Several space craft (TIE-fighters) intersect in the sky, two of them hitting each other emulating the motion bounce illusion from experiment 3. The crafts are identical to the one in experiment 2, building upon the implementation of moving objects emitting sound. The camera continues up a flight of stairs seeing more crafts, one moving to the right, drawing the camera to a large space ship (A Star Destroyer, also from star wars), emitting sounds at a great distance. An explosion occurs in the distance at this large space craft that then also shoots directly at the camera cutting the video abruptly. The sounds of these explosions are created on the basis of the manipulation done on sounds for the first experiment, and from the explosion in the 4th experiment.

The scene in its entirety was meant to show an initial stage of a planetary invasion by the imperial navy from the Star Wars saga creating the small desired narrative for the video.

Design & implementation

The first step was to record the raw footage, which had to be done with future camera tracking in mind. The camera was pointed to the sky twice, each time carefully and as smoothly as possible to make the camera tracking easier, as rapid movements could cause complications.

Camera tracking, or match-moving as it is sometimes called, was then applied to the two shots filming the sky, using software to map the movements of the camera by analyzing the movements of objects filmed – specifically it analyzes the path of high-contrast spots and extrapolates the camera movement from their angular velocities, or pixels/s. This analysis was done with Bojou 5.0 (see snapshot of tracking in Figure 29), and the solved camera outputs were exported for use in 3DSmax. (see Appendix D: Camera Tracking in Boujou. for an additional illustration of the tracking procedure)

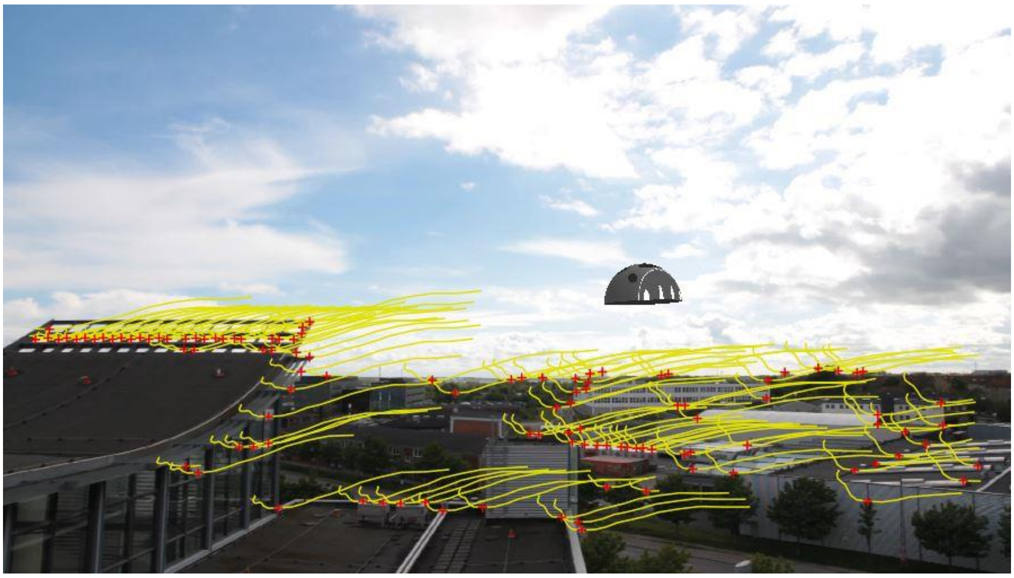


Figure 29 - Camera tracking / match-moving for Experiment 5. The grey ladybug being a debug object for testing purposes

As mentioned, some 3D models were recycled from experiment 2 (TIE fighters). – again from their familiarity but for this experiment also their affiliation to other space craft. The very large craft known as a Star Destroyer from the star wars saga were used to experiment with the sense of scale and to compliment the smaller craft mentioned before. The model of the star destroyer obtained through the scifi3d portal (Johansson, 2002), as were the model for the Tie Fighter (Bassett, 2000). Additionally, a model of an imperial shuttle (Allamandri, 1999) were added from the same database, used briefly as a transitional pan to the star destroyer.

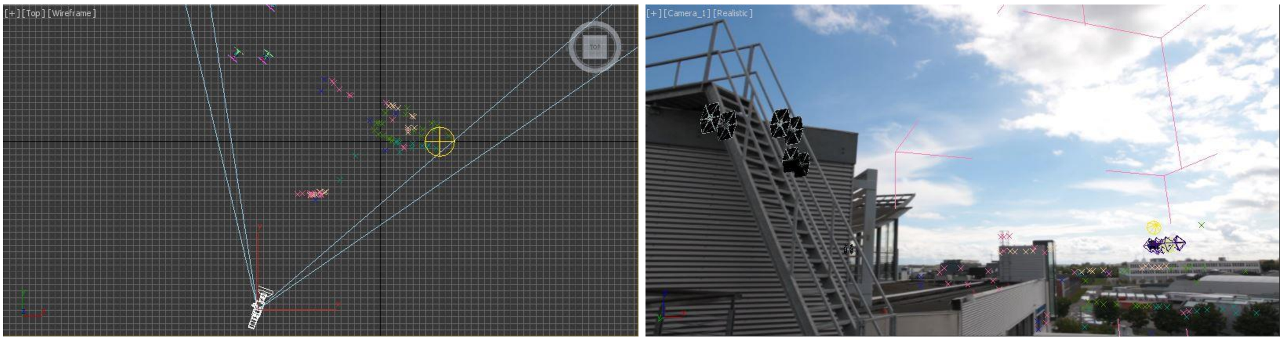


Figure 30 - 3DS MAX scene of Experiment 5 (first shot). The small crosses indicate extrapolated tracking marks.

These models were lit matching the conditions of the footage and animated in 3D Studio Max, where the models were placed according to the extrapolated positions of the tracking spots performed in the previous step in Boujou 5.0 (see Figure 30 for preview in 3DSmax). The scene was rendered out for final compositing in Adobe After Effects CS6 which also added motion blur. The compositing involved matching color and brightness of the models, and the addition of clouds in front of distant objects. Some objects received a gradient of grey from the top simulating low cloud cover on medium-far distance objects, and all elements received matching camera grain and blurring (see compositing in Figure 31). Two laser shots were added firing from the star destroyer, created with the "CC Beam" effect in After Effects, followed by distant explosion and one on top of the camera. These had to be tracked onto the terrain, which was done with single point tracking. Following the visual compositing, a light color correction was performed to raise the gamma and have the scene in less hard contrast.

Figure 31 – Compositing Experiment 5, frame 377-

The soundscape was created in Sony Vegas 12c. The many animated elements each needed to have appropriate sounds and manipulations distinguishing them from each other. "Sweetening" effects of rushing air were added as crafts passed closely, and the Doppler Effect were created for crafts moving to and away from the camera. Pitch, amplitude and stereo pan was created as envelopes for all elements (see Figure 32), along with low-pass filter to simulate being outdoors (as described in experiment 1) and a hint of reverb because of the many building facades nearby.

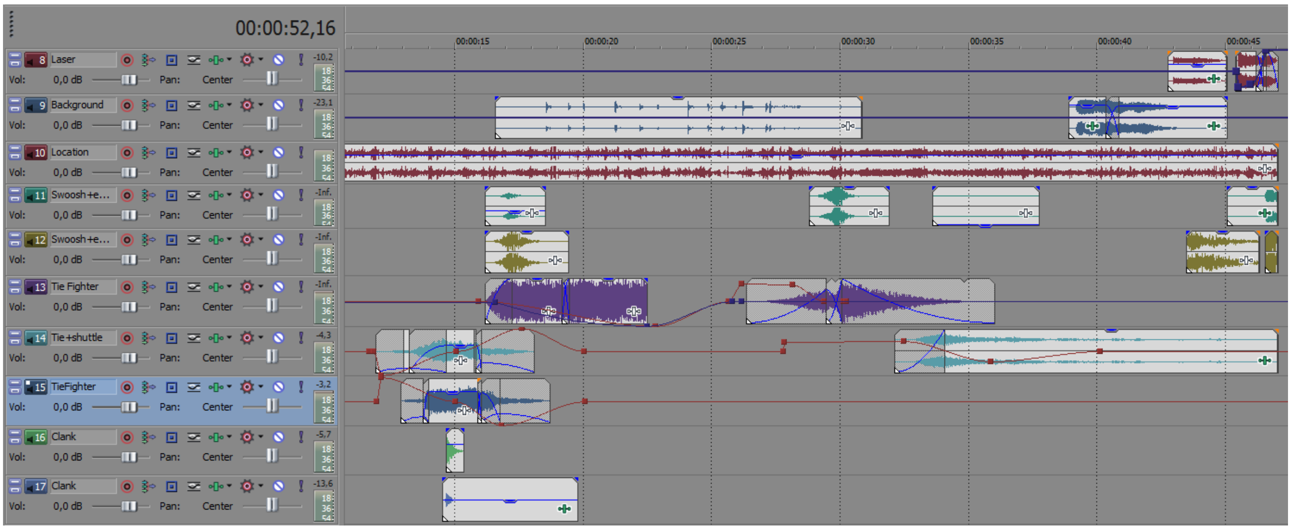


Figure 32 - Sound compositing for Experiment 5. Horizontal colored lines represent amplitude and stereo pan envelopes

The sound of the TIE fighters were the same as in Experiment 2, altered in pitch between each craft and their direction. The rushing of air as they passed close was from the Motion Pulse soundpack from Video Copilot, labelled as "swooshes". A recording of a horn at great distance was added to the Star

Destroyer as it comes into frame, originally sampled from the War of the Worlds movie from 2005. This was to have the massive space ship appear more menacing and the horn should be understood as a warning. The explosions were of same origin as those in Experiment 4, with debris close and at high amplitude, with the far away explosion being a snippet of a simulated recording of a battlefield.

The video sequence was 47 seconds long, running at 25 frames per second in 1920x1080 resolution, and the audio sampled in 48 kHz. A list of stills from the experiment video is available in Appendix E: Stills from Experiment 5 video; Chronological order

Testing

The video was shown to 25 participants in an online questionnaire after which there were asked to rate the following aspects of the video:

- Overall Quality
- Visual Quality
- Sound Quality
- Realism

As described, this test did not seek to prove or disprove any hypotheses, as instead it attempted to combine the previous experiments into one video of high quality, which was why these elements were tested..

Results

Both visual and sound quality scored high ratings, with visuals having an average of 3,8/5 and sound having an average of 4,2/5 with distribution of ratings seen below in Figure 33:

Figure 33 - Assessed Visual and Sound quality of Experiment 5 on a scale from 1-5, 1 being poorest and 5 being best.

The ratings of sound quality was significantly different from the ratings of visual quality by a p-value of 0,03 ($p<0,05$), indicating that sound quality was better than the visual quality, which was still moderately high.

Participants were also asked to assess the videos overall quality, resulting in an average answer of 3,84/5 as seen below, showing that participants generally thought the video and its elements to be of good quality with a distribution as seen below in Figure 34.

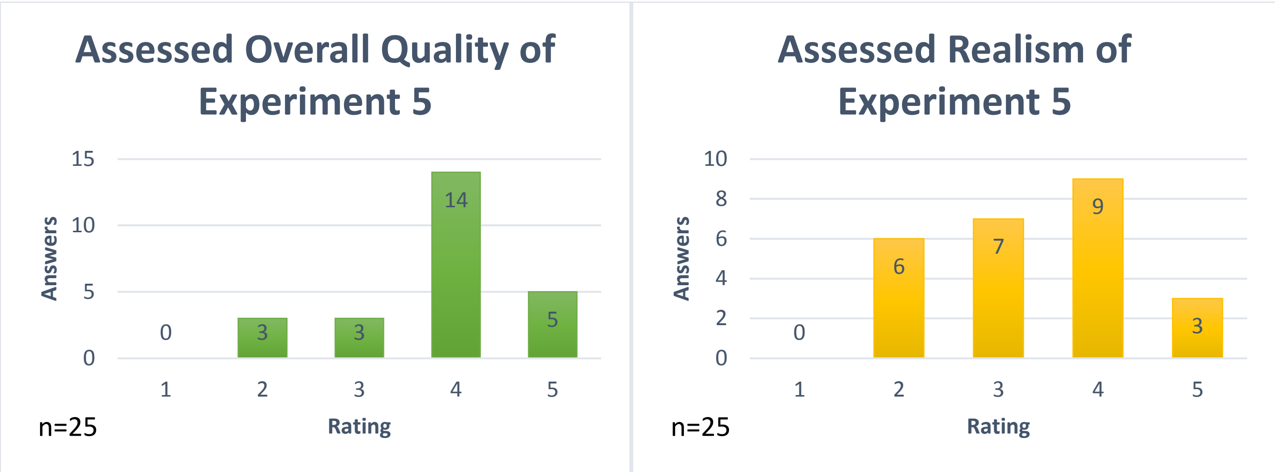


Figure 34 - Assessed Overall Quality and Realism of Experiment 5 on a scale from 1-5, 1 being poorest and 5 being best.

Participants were also asked to rate the video's realism, as in if it looked and sounded realistic. Here the ratings were more mixed with an average of 3,36/5 and a distribution as seen above in Figure 34. This also reflects the creator's assessment, as the last percentages of realism were difficult to achieve, which potentially broke "the illusion of realism" or made video appearing in some manner of "uncanny valley".

(All participant ratings available in Appendix F: All ratings of Experiment 5)

Overall, the video in Experiment 5 received positive ratings in the four categories tested, as well as in comments to the video in the questionnaire. With these results, the experiment is called a success, with room for improvement.

Discussion of Investigation

The results and experiments in this investigation can in some regards be considered as invalid due to the tested aspects not being 100% isolated due to the amount of noise present in the tested perceptual modalities. Part of the goal of this investigation was to carry theories into rich media, which inherently held more noise, which can in many way skew the results.

From the onset of the investigation another goal was to achieve high production quality. In the implementation of theories into video, a lot of synthesized elements were added to real recorded video. These synthesized elements were not always successfully integrated into the test videos, making them stand out and look unrealistic and causing some elements to appear "fake" likely affecting results.

Examples of elements being of lower quality includes low-detail lightning-strikes in experiment 1, simplified interaction with water surface in experiment 2, "flat" explosion in experiment 4 and errors in camera tracking in experiment 5.

The data collected for the experiments were of low volume, rendering conclusions uncertain. More results would need to be acquired to effectively conclude on the experiments in the investigation.

Likewise, there is a potential bias in that the majority of participants followed the same study as the author, making them hypercritical of work in tests or easily excited as they perhaps knew the conductor personally.

The difference of results from the personally conducted tests and autonomous online tests, may invalidate results and conclusions in experiment 1 and 2. The cause of these are as mentioned unknown, and any further research would warrant a deeper investigation into what trigger this disparity.

Due to the desired volume of tests and products desired, each experiment avoided going into detail on many areas. This means that the experiments have weaker results, tests and products than if a single experiment had been chosen. As such, volume of tests and products were favored over potential quality, raising questions if this was the best choice. However it is of this investigation's opinion that broad knowledge of several areas is as valuable as specific knowledge of a single area.

Conclusion of Investigation

Experiments in this investigation suggest that audio is unsurprisingly important for the creation of video with synthetic elements. As sound had a large influence in the perception of videos created, it proved capable of changing the perception of numerous visual events – even creating new perceptions of events that did not take place.

The importance of sound in film, is not a new discovery, as it has been part of video production for as long as sound and image has existed together. While this investigation may not the first to arrive at its conclusions, it helps reaffirm theories established in the past, such as synchresis and the audience' general perception.

This investigation also proves that it is possible to take isolated and simple audiovisual theories into practical application as video, where results show that these theories were often still in effect despite the complex and noisy nature of the rich video products.

Overall, the investigation successfully implemented and tested crossmodal interaction in video, carrying audiovisual theories into new application, providing additional results from experiments done with high visual- and auditory quality following new interpretation.

Bibliography

Allamandri, F., 1999. *3D Model: Imperial Lambda Class Shuttle (Tydirium).* s.l.:SciFi3D.

Bassett, J., 2000. *3D Model: Tie Figher SE.* s.l.:SciFi3D.

Biocca, F., Kim, J. & Choi, Y., 2001. Visual Touch In Virtual An Exploratory Study of Presence, Multimodal Interfaces, and Cross-Modal Sensory Illusions. *Presence*, 10(3).

Carlesa, J. L., Barrioa, I. L. & Luciob, J. V. d., 1999. Sound infuence on landscape values. *Landscape and Urban Planning,* Volume 43, pp. 191-200.

Carlsson, S. E., 1994. *When Picture and Sound Merge.* [Online]
Available at: http://filmsound.org/when_picture_and_sound_merge/

Chion, M., 1994. *Audio-Vision - Sound on screen.* s.l.:Columbia University Press.

Cohen, A. J., MacMillan, K. & Drew, &. R., 2006. *The Role of Music Sound Effects & Speech on Absorptionin a Film :,* Charlottetown: University of Prince Edward Island.

Cohen, S. K. M. a. A. J., 1988. Effects of Musical Soundtracks on Attitudes toward Animated Geometric Figures. *Music Perception: An Interdisciplinary Journal,* pp. 95-112.

Conway, A. R. A., Cowan, N. & Bunting, M. F., 2001. The cocktail party phenomenon revisited: The importance of working memory capacity. 8(2).

Gibbs, K., 2013. *Projectiles and air resistance.* [Online]
Available at: http://www.schoolphysics.co.uk/age16-19/Mechanics/Kinematics/text/Projectiles_and_air_resistance/index.html

Hirsh, B. J. a. I. J., 1972. Visual and auditory perception —an essay of comparison. *Human communication: A unified view,* pp. 283-340.

J. M. Brown, K. L. A. C. A. F. a. C. C., 1998. Visual influences on auditory distance perception. *J. Acoust. Soc. Am.,* p. 1798.

Jekosch, U., 2005. *Assigning Meaning to Sounds – Semiotics in the Context of Product-Sound Design,* New York: School of Architecture, Rensselaer Polytechnic Institute.

Johansson, F., 2002. *3D model: Imperial Star Destroyer: Four.* s.l.:SciFi3D.

Joly, A., Montard, N. & Buttin, M., 2001. *Audio-Visual Quality and Interactions between Television Audio and Video,* Villeurbanne: 'TDF-C2R Broadcasting and Wireless Research Center of Télédiffusion de France.

Labor, U. S. D. o., 2014. *Appendix I:A. Physics of Sound.* [Online]
Available at: https://www.osha.gov/dts/osta/otm/noise/health_effects/physics.html
[Accessed 14 March 2014].

Para, A. K. a. S. v. d., 1999. *Auditory-Visual Interaction: From Fundamental Research in Cognitive Psychology to (possible) Applications.* San Jose, s.n.

Rinzler, J. W., 2001. *The Sounds of Star Wars.* s.l.:Chronicle Books.

Sanabria, D., Correa, Á. & Lupiáñez, J., 2004. *Is there a timing synergy during multi-finger production of quick,* Verlag: Springer.

Sekuler R, S. A. R. L., 1997. Sound alters visual motion perception. *Nature*, 385(308).

Shimojo, K. W. a. S., 2001. *WHEN SOUND AFFECTS VISION: Effects of Auditory Grouping on Visual Motion Perception,* s.l.: California Institute of Technology-Pasadena.

Sonnenschein, D., 2001. Sound Design: The Expressive Power of Music, Voice, and Sound Effects in Cinema. In: s.l.:Michael Wiese Productions, pp. 1-20.

Sugita, Y. & Suzuki, Y., 2003. Implicit estimation of. *Nature,* p. 911.

Vavrek, R. J. et al., 2014. *The Science of Thunder.* [Online]
Available at: http://www.lightningsafety.com/nlsi_info/thunder2.html

Welsh, R. B. & Warren, D. H., 1980. Immediate perceptual response to intersensory discrepancy.. *Psychological Bulletin*, 88(3).

Wolfe, J. M., Kluender, K. R. & Levi, D. M., 2009. Hearing in the Environment. In: *Sensation & Perception.* Sunderland: Sinauer Associates inc., pp. 250-273.

wookiepedia, 2014. *TIE/LN starfighter.* [Online]
Available at: http://starwars.wikia.com/wiki/TIE/LN_starfighter

Zahorik, P., 2002. Accessing auditory distance perception using virtual acoustics. *J Acoust Soc Am 111,* pp. 1832-1846.

Appendix

Appendix A: Experiment 1 Conducted Interview Structure

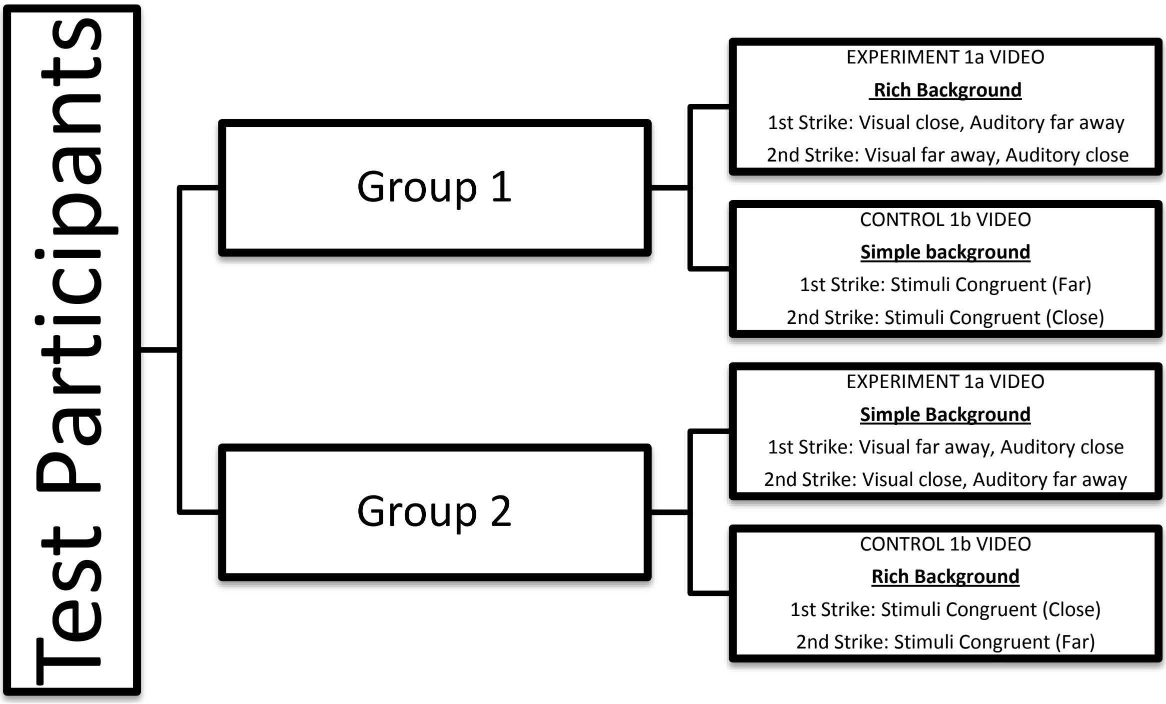


Figure 35 - Experiment 1 Structure

Appendix B: Picture of Tennis ball used for Experiment 3

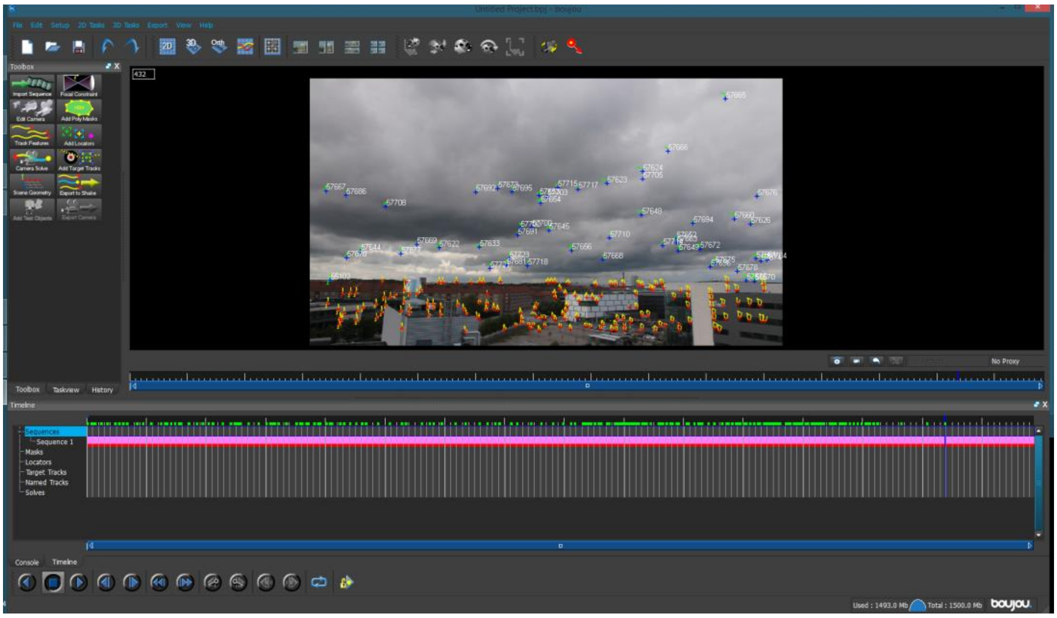
Appendix C: Music and Ambience

Music and video is a long-standing tradition that started with the earliest of silent movies then accompanied by a musician in the theatres often playing a piano. The reason that these media go well together is that they complement each other with the type of information they present. Where music is able to describe emotional information, it cannot describe specific events in detail, while video can do both (Cohen, 1988). Because of this, music is used to emphasize emotion, creating potentially stronger emotional response with the audience.

Cohen establishes in 1988 that *"The temporal component of music may affect interpretation of film through congruence with the action or pattern of motion of the film."* Her studies suggest that music changes our attentional strategy towards a film, and subsequently our understanding and decoding of it. This is recognized especially when interpreting characters in film, where the music plays a large part in creating congruence between the character and its nature and intentions. (Cohen, 1988)

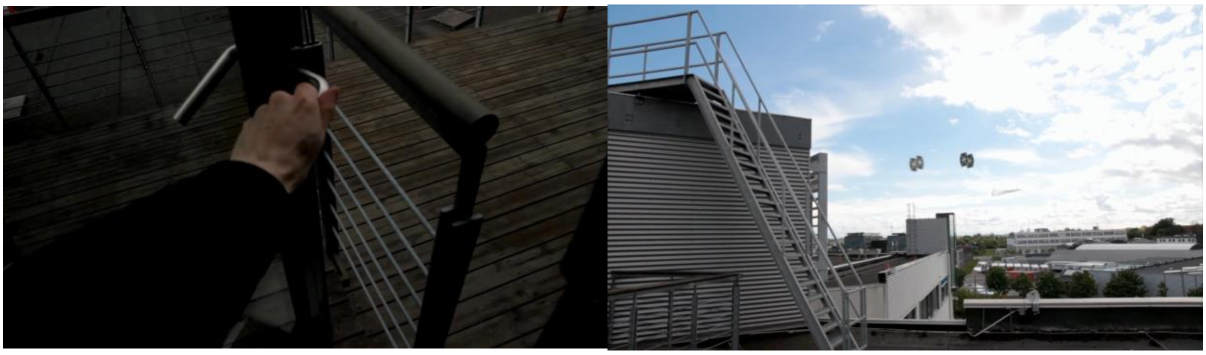
The emotional influence of music can provide a context for an otherwise visually neutral scene. Likewise, a shared rhythm between visual motion and music can be used to draw attention to the motion and movements. As such, music helps video by contributing with emotional information and aiding the storytelling (Cohen, et al., 2006). In a study by Annabel Cohen, participants were exposed to a one minute video with and without various pieces of music, resulting in efficient absorption in the material even with this short exposure. They suggest that music alone may hold a privileged path to absorption and that this absorption occurs quickly.

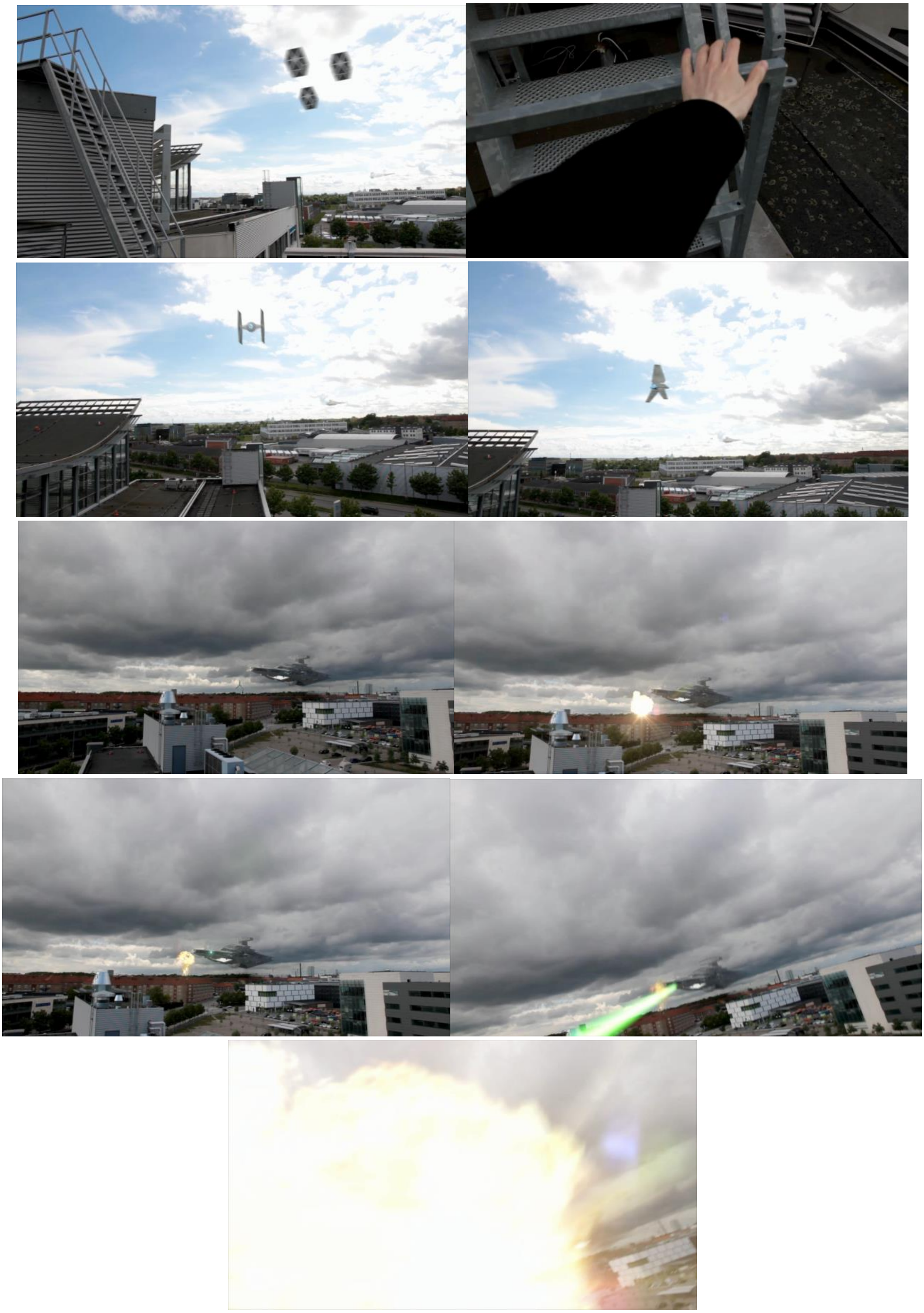
Appendix D: Camera Tracking in Boujou.



Note; all the blue/green dots are selected tracking marks, meant for deletion as the clouds were not considered stationary for tracking.

Appendix E: Stills from Experiment 5 video; Chronological order





Appendix F: All ratings of Experiment 5 video

	Please rate the overall quality of the video you just saw	Please rate the VISUAL quality of the video you just saw	Please rate the SOUND quality of the video you just saw	Please rate how realistic this video looked and sounded
	4	4	4	4
	3	3	4	2
	5	4	5	4
	4	4	3	2
	2	3	4	2
	5	5	5	5
	2	4	3	3
	3	2	4	2
	4	3	4	3
	4	4	5	2
	4	4	4	3
	4	4	4	3
	4	5	5	4
	4	4	3	3
	3	3	4	4
	5	4	4	4
	4	3	5	4
	5	5	5	4
	2	2	2	2
	4	4	5	3
	4	5	4	4
	4	4	5	4
	4	4	5	3
	4	4	4	5
	5	4	5	5
AVG:	3,84	3,8	4,2	3,36
Summary				
1	0	0	0	0
2	3	2	1	6
3	3	5	3	7
4	14	14	11	9
5	5	4	10	3