Influence of Surround Sound on Visual Fixations During Voluntary Video Viewing

Manolis Fragkiadakis
Graduation Thesis, August 2016
Media Technology MSc program, Leiden University
Thesis advisors: Maarten H. Lamers and Bernhard Hommel
mfragiadakism@hotmail.com

***Abstract*—Studies have shown that gaze behavior is affected by visual features such as motion, brightness and contrast. In addition, it has been assumed that monophonic sound content can also affect the allocation of gaze. However, surround sound has not been studied as a possible factor that can influence eye behavior in dynamic scenes. In this study we investigate the influence of surround sound on eye movements during voluntary video viewing. We recorded the eye movements and in particular the fixations of 21 participants during videos of outdoor scenes with monophonic and surround uncorrelated to the content of the videos sound. The results showed that, in general, fixations with surround sound were clustered at approximately 130 pixels away from the ones with monophonic sound. Furthermore, in moments when there were moving objects appearing in a scene, the sound condition did not modulate the allocation of gaze. Finally, the locations of the clusters of multiple participants are significantly dependent on the location of the active speaker. While surround sound influences gaze allocation, motion seems to be a visual feature that interferes with this impact. However, limitations of the eye tracking device and the spatialization technique used in the experiments prevent us from drawing further tangible conclusions.**

***Index Terms*—eye-movements, attention, video, surround sound, audio-visual integration, eye-tracking experiment, visual attention models.**

I. Introduction

THE scientific topic of visual attention has been thoroughly studied over the years [1]. Attention, the process of focusing on one aspect of the environmental information while ignoring others, has been researched by both behavioral psychologists and neuro-scientists.

Visual attention is a concept that comprises of multiple meanings including selection, concentration, focusing, vigilance and the like. However, in this research we will concentrate on the domain of selection. Research has shown that while humans receive a large amount of information from their surroundings and the environment, only portions of it is being processed, while the rest is being ignored [2] [1]. The reason for this according to Lennie, is the fact that brain's capacity is limited, thus the energy to efficiently activate all the neurons is limited [3].

Moreover, studies have shown that visual attention in highly linked with eye movements. Experiments by Rizzolatti et al. showed that there is a correlation between the oculomotor system and visual attention [4]. Recent neurophysiological experiments strengthened this theory [5]. Although other studies indicate that these processes are greatly separated [6] there is a widespread assumption that attention is highly correlated with eye movements.

As a result, eye movements can be considered as "the real time index of visual attention and cognition" [7]. There are four basic types of eye movements: saccades, smooth pursuit movements, vergence movements and vestibulo-ocular movements. Saccades are rapid, ballistic eye shifts that abruptly change the point of fixation. As described by Henderson, during active real-world scene perception "attention is typically directed to the fixated location and the location to be fixated next" [8].

A. Direction of attention

In the study of attention there is a major distinction between goal-directed factors (also called endogenous control) and stimulus-driven factors (also called exogenous control). The first ones refer to factors that focus attention on cognitively relevant features of the environment while the last ones refer to the unintentional attention that is being captured by external, stimulus features such as luminance, color and contrast [9][7]. This means, that in a freely, voluntary state attention should be persistent across multiple viewers since they share the same stimuli. In addition, endogenous control should result in less coordinated attention as the individual cognitive state differs and is less predictable[10][2][11].

B. Visual saliency

The aforementioned features led scientists to develop several computational models to predict attentional allocation in static and dynamic scenes [12][13][14]. These models according to Koch and Ullman, assume that discrete visual features emerge and involuntary attract attention [15]. Additionally, these models were used to create saliency maps: topographically arranged maps to predict "the conspicuity of specific locations and their likelihood of attracting attention" [7]. However, most of the studies investigating the link between gaze allocation and visual saliency excluded a transient feature: motion. Experiments have shown that the only moment where gaze of multiple viewers is clustered at the same location can be predicted by motion, since the associated motion of an appearing new object into a scene, can exogenously attract attention [7][16].

C. Audio integration in visual attention

In general, sound has rarely been considered a possible feature that can drive and influence attention in dynamic scenes. Most of the studies in dynamic scene allocation and eye movements have excluded the original soundtrack[1] (sometimes intentionally) and forced participants to watch soundless movies [12][17][7]. However, in 2012 Coutrot et al. [18] researched on "the impact of non-spatial sound on the eye movements of observers watching videos". They concluded that sound indeed has an impact on eye position and fixation but it is not constant across time. By recording the eye movements of viewers in videos with (AudioVisual condition) and without sound (Visual condition) they concluded that the gaze allocations differ significantly in each condition and that participants in the AudioVisual condition make larger saccades. Specifically, they have found that the influence of sound in a scene appears after the 25th frame, as in "the beginning of a scene exploration, the influence of sound is outweighed by visual information" [18].

Later on, Coutrot and Guyader showed that in conversation scenes, faces are the highest gaze attractors and when viewers hear the original soundtrack of the scene, they are able to "follow the speech turn-taking more closely" [19]. These results also confirm experiments by Song et al. where it was shown that different types of sounds influence gaze differently in videos and that human voice has the greatest effect [20].

Generally, the visual enhancement by sound has also been investigated by Zou, Muller and Shi where they concluded that "spatially uninformative sounds facilitated the orientation of ocular scanning away from already scanned display regions not containing a target and enhanced search performance" [21]. Furthermore, it was found that the accuracy and speed of eye movements during detection tasks was improved when audio-visual stimuli were presented to the subjects compared to a mere auditory or visual stimulus [22][23][24].

However, none of these studies (apart from [21]) have used surround sound in their experiments. Quigley et al. [25] studied the influence of spatially localized (left, right, up, down) sounds in static natural images and it was shown that eye movements were spatially biased towards the sound sources.

Nevertheless, dynamic scenes have not been used in this context. In order to fully understand which properties of surround sound affect gaze allocation in dynamic scenes (as has been previously studied with images), we first need to investigate whether this sound condition has any effect at all.

D. Surround sound and visual attention

In this research we investigate the influence of azimuthal surround sound on eye movements during voluntary video viewing. Does the location of a sound source modulate attention in videos? Additionally, are the eye movements of multiple participants clustered in the same location based on the sound condition?

[1]The term "soundtrack" refers to all the sounds (music, dialogues etc.) that are accompanying and synchronized with a video.

The results from the experiments of Quigley et al. [25] showed that gaze is biased towards the location of the image corresponding to the sound source. Additionally, experiments by Song et al. suggest that "the sound source in the video seems to attract attention" [26]. Considering all the above we hypothesize that the locations of the clusters in the Surround condition will be significantly correlated with the location of the sound source playing, as assumed by the previous studies. Although monophonic sound has not been compared with surround sound in dynamic scenes, if the fixations in the latter condition are biased towards the sound source then the allocation of gaze will be substantially different between the two sound conditions.

Boltz showed that audio and music in films can highly influence the general emotional impact and interpretation [27]. We live at a time where film makers use surround sound systems more and more in order to make an experience more immersive. We believe that the current study can provide the necessary information in order to use these systems more efficiently and utilize the location of a sound source as an attention attractor in relative screen areas.

Additionally, this study can provide groundwork to further explore the features of surround sound in dynamic scenes and lead to saliency models that consider this sound condition as a gaze steering factor. Positive results will provide the stepping stone to extend surround auditory stimuli beyond static content.

II. Methodology

A. Participants

21 people were asked to participate in the research: 13 male and 8 female with ages ranging from 22 to 55 years (M = 28). All had normal or corrected to normal vision and reported normal hearing. Additionally, they were informed about the general purpose of the research and gave their consent to participate. This study was approved by the local ethics committee.

B. Stimuli

The 10 videos used in the experiments were of natural outdoor scenes sourced from publicly accessible repositories and included parts from advertisements, documentaries and movie scenes. Each video had a 1440 x 1080 pixel resolution at 30 frames per second and lasted 20 seconds. As a whole, video sequences last approximately 7 minutes. All videos were encoded using Adobe's Premiere Pro CS6 in a H.264 video compression standard in order to provide good video quality at lower bit rates. The videos were chosen carefully in order to avoid any distinctive objects at their center but having an interesting spatial context. Studies have shown that fixations are biased towards the center of the screen [11][28]. While this center bias is usually apparent at the beginning of a new scene [29] we chose videos that have non-substantial information presented at this screen area. Table 1 presents a concise overview of the selected videos.

For the auditory stimuli, 10 sounds were used to be presented in the videos (Table 2). The content of these was

TABLE I: Video Descriptions

ID	Title	Description
1	Street	street with moving cars and people
2	Cosmos	earth and sun with high contrast
3	Creek	water creek from left to right
4	Islands	sunset with motion in forward
5	Coast	coast during sunset
6	Space	stars moving towards the viewer
7	Stream	water stream from right to left
8	Running	dock with people running
9	Sidewalk	blurry sidewalk with people walking
10	Waterfalls	waterfalls in fast motion

TABLE II: Descriptions of the sounds used in the experiments

ID	Description
1	cleaning and whipping carpet
2	electric wheelchair pull up
3	ice tinkling in full glass
4	falling large tin can
5	opening leather bound wood jewelry box
6	manual hedge trimmer
7	push metal trash can
8	remove lid from metal garbage can
9	scrubbing rug with brush
10	watering plants with watering can

unrelated to the videos in order to avoid any semantic bias and consisted of household and office appliances sounds lasting 4 seconds respectively (c.f. [20]). As sounds related to the content of the videos would have directed gaze into salient prominent features, we chose unassociated auditory information. The auditory stimulus was considered appropriate for our study, as none of the videos contained indoor shoots or houses in particular. Additionally, it might be interesting to investigate whether unrelated sounds will lead gaze to transpose the salient feature's threshold. Furthermore, each video soundtrack consisted of a randomized sequence of 5 of these sounds, each playing in a discrete sound channel without empty, soundless moments in between. Thus each video had one corresponding soundtrack in 5.1 surround mode and one in mono. The order of the sounds was the same for the mono and the surround soundtrack. Each sound was manually converted in 5.1 surround standard using Audacity (open source digital audio editor) in order to create the appropriate "spatalization" or used in mono when needed. All sounds were normalized to approximately equal amplitude. All sound channels (front left, front right, rear left, rear right) were mono apart from the center "speaker" that was in stereo. Finally, the order of the sound channel that was playing in each video respectively was the same between the subjects.

C. Apparatus

All experiments took place in a dimly-lit room at Leiden University dedicated for this study. Participants were seated 60 cm away from a 22 inch LED monitor in a resolution of 1440 x 1080 pixels and at 60 Hz refresh rate with its center reaching approximately the eye level of the subjects. In order to stabilize the head, a chin rest was used.

The audio stimuli were presented via the Razer Kraken Pro 7.1 surround headphones using Razer's corresponding audio engine. Additionally, the volume was chosen by each subject before the experiment to match a comfortable level. While a physical surround speaker setup would have been more appropriate for this study it was not possible to present the audio stimuli through the software used in the experiments. Thus a headphone solution was used.

TABLE III: Video sequences of 5 participants and the respective sound conditions (M for Mono and S for Surround)

Participant	Video sequences	Soundtrack condition
1	8,9,2,5,6,3,10,1,4,7	S-M-S-M-M-S-S-S-M-M
2	2,7,5,9,8,1,6,10,3,4	M-S-S-S-M-M-S-M-M-S
3	9,1,8,7,5,3,2,10,6,4	M-M-S-S-S-M-M-M-S-S
4	8,9,2,1,10,6,4,5,3,7	M-S-S-S-S-M-M-M-S-M
5	1,10,9,8,3,4,6,7,5,2	M-S-S-M-M-M-S-S-S-M

Eye movements were recorded using an Eyetribe eye tracker at a sampling rate of 60 Hz. The recorded data consisted of the binocular gaze data (x/y screen coordinates) and pupil diameter in mm. Before each experiment a calibration phase was held which consisted of a circular target displayed at 9 different positions of the screen on a blank background during 2 seconds each. Once the subject had looked at all the targets the calibration process had been completed.

D. Procedure

The experiment was designed using OGAMA open source software. The software allows the recording and the analysis of eye tracking data from slide show eye tracking experiments in parallel. Each experiment consisted of a Monophonic (Mo) and a Surround (Su) condition. Before each video, a fixation cross was presented in the center of the screen for 2 seconds in a black background. After that, each video was played in full screen. Figure 1 presents the time course of an experimental trial.

Subjects had to look involuntary at 10 videos. To avoid any order effect and bias, each experiment had a randomized sequence of videos (Table 3). Five of them played with mono sound and five in surround mode. Each subject looked at each video only in one condition. Thus each video was seen in the mono condition by 11 subjects and in the surround condition by the other 10 subjects.

After the eye tracking experiment, participants had to listen to 5 sounds (each one playing from only one speaker at a time with 2 seconds pause in between) and fill in a questionnaire regarding the location of the speaker. The reason of this task was to determine whether the spatialization technique and engine used in the experiments could create the illusion of an appropriate surround speaker setup.

III. Analysis

A. Data

The recorded fixation data were extracted from Ogama's database module containing all the raw screen coordinates and eye movements. The eye tracking device captured 60 eye positions per second, thus 2 positions per frame. To match the frame rate of the videos we used the median position of these 2 coordinates and only when a fixation occurred. All data containing blinks or saccades were discarded from the eye

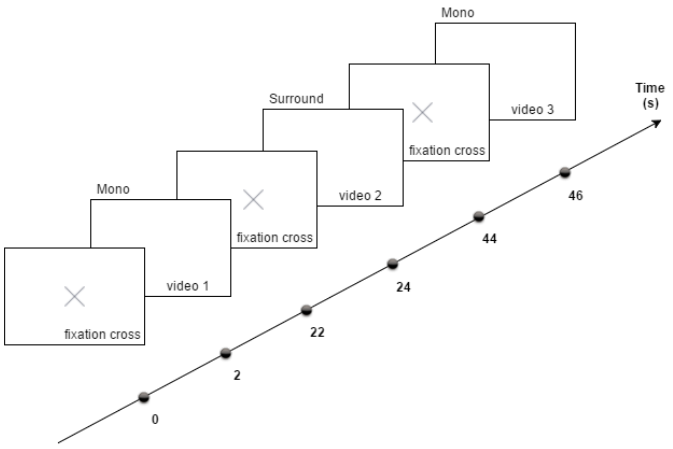


Fig. 1: Procedure of the eye tracking experiment. Each of the 20 trials lasted for 20 seconds and the sound playing was either in mono or in surround mode. Before each trial a fixation cross was presented.

TABLE IV: Mean positions of the clusters for the first 20 frames of Video 1 in each sound condition

Frame	Mean position in Mono		Mean position in Surround	
1	0	0	973	506
2	779	341	0	0
3	0	0	952.5	527.25
4	0	0	953.667	513.333
5	764.5	235.5	953.5	538
6	792.329	316.899	948.506	408.264
7	783.429	313.571	949.305	409.412
8	764.5	235.5	930.358	530.737
9	792.329	316.899	961.853	415.544
10	781.945	271.296	957.745	447.758
11	783.429	313.571	760.364	484.545
12	685.5	274.375	741.4	498.3
13	688.75	250.375	743.111	526.222
14	688.375	251.625	727	544
15	688	251.125	721.778	533.889
16	766.111	308.333	714	525.111
17	736.75	322.75	705.778	522.778
18	724.875	332.625	634.667	516.556
19	656.625	386.5	603	519.222
20	653.125	382.75	599.222	522.778

position analysis since we are only interested in the positions of the fixations. Using Ogama's fixation module we verified the time and position of the fixations using 10 pixels as the maximum distance that a point may vary from the average fixation point and still be considered part of the fixation.

Additionally, we separated the data of all eye movements, per video, into two different sets depending on the sound condition. Thus, each video had a corresponding set of all the subjects that watched it in Mono and in Surround condition respectively.

B. Clustering

In order to find whether and where the eye positions of all participants were clustered in each video and sound condition, we used the Computational and Algorithmic Representation and Processing of Eye-movements (CARPE) software [7]. The software uses a Gaussian Mixture Model (GMM) to "classify moments of tight and loose clusters of eye movements". GMM, as a soft clustering method, assigns a score to a data point for each cluster which indicates its strength. In general, the model in question has been previously used as a clustering method[30] and is considered to be a better approach compared to k-means clustering as it accommodates clusters that have different sizes and correlation structures within them [7]. For each frame we extracted the mean position (in pixel coordinates) of the clusters found and their respective weight using one to eight clustering kernels and spherical covariance type (Table 4). When multiple clusters were found we chose the one with the highest weight.

C. Metrics

In order to estimate the difference of the clusters between the Mono and the Surround condition we used a metric called dispersion. The dispersion D for a frame f is defined as:

$$D(f) = \sqrt{|y_S - y_M|^2 + |x_S - x_M|^2} \quad (1)$$

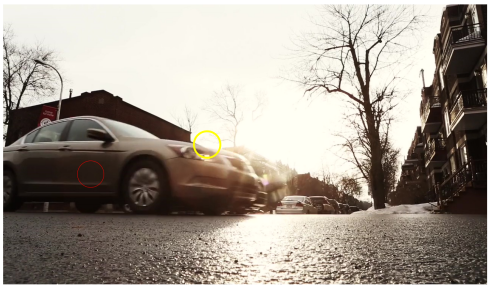
where, x and y are the pixel coordinates of the Mono(M) and Surround(S) clusters respectively. Dispersion is essentially the euclidean distance of the positions of the two clusters for a given frame. If dispersion is relatively small, then the clusters are close to each other, while when dispersion is high, the two clusters are far apart. First, we calculated the mean dispersion in each video over all frames (global analysis). Additionally, we studied the development of dispersion across the frames and compared the results between all the videos (temporal analysis). If the sound condition influences where fixations are located, then the dispersion should be high across all the frames and for all videos. However, this metric has some limitations. First and foremost, it does not show which condition influences most the scattering of the clusters. Additionally, it does not provide any information about the evolution of the clusters across the frames. In the analysis, we computed the dispersion for each frame for all 10 videos. We also plotted the positions of the clusters over each video in the two sound conditions in order to understand the areas of interest.

IV. Results

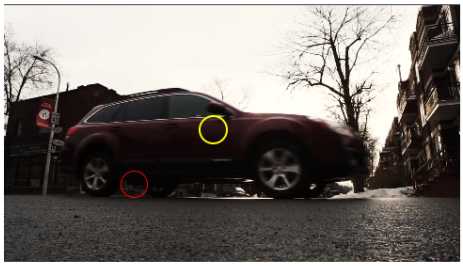
The objective of this study is to investigate how surround sound influences the allocation of eye gaze in comparison to monophonic sound during voluntary video viewing. As a result, we compared the recorded eye fixations in each video, between the Mono and Surround sound condition respectively (comparative dispersion) and between the fixations of participants within each sound condition (internal dispersion). Firstly, we analyzed how dispersion between the two conditional clusters is developed over time. Then, we focused on the location of the sound source playing in each video and its respective influence on the positions of the clusters.

A. Comparative Dispersion

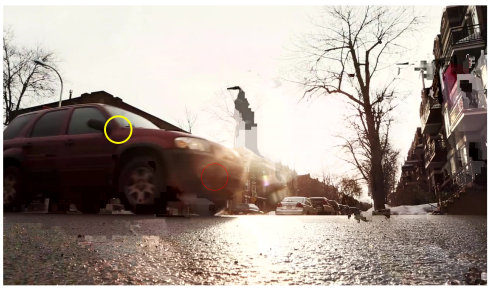
1) Global analysis: In order to understand whether Surround sound influences in general the positions of the eye



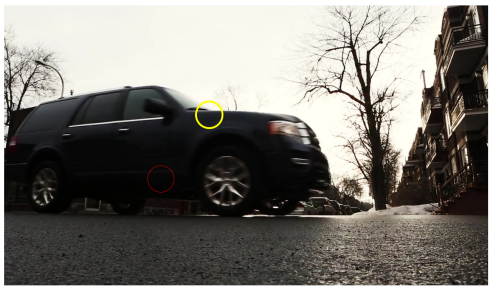
(a) Frame 55



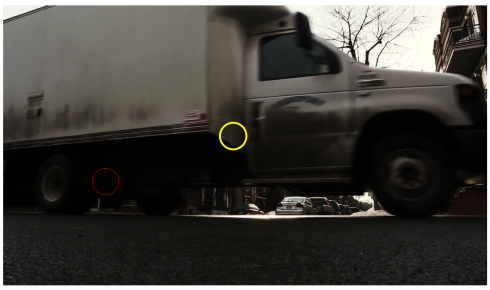
(b) Frame 150



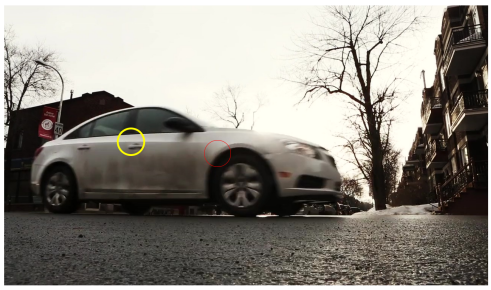
(c) Frame 246



(d) Frame 354



(e) Frame 443



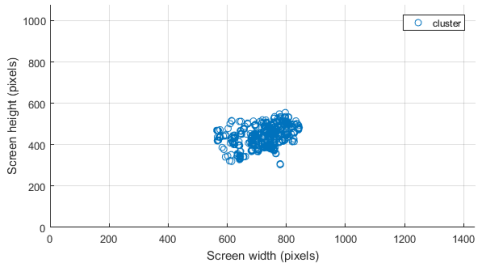
(f) Frame 547

Fig. 2: Frames that correspond to the local maxima of dispersion in Video 1. Yellow circles present the position of the clusters in Surround condition, while in red are the positions of the clusters in Mono condition.

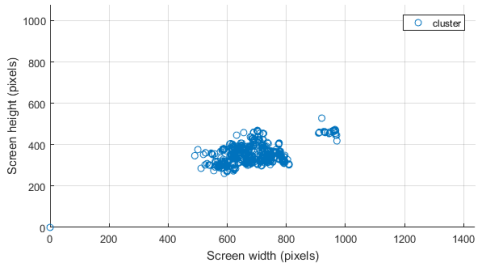
fixations compared to Mono sound, we calculated the mean dispersion for all videos and within all frames. Generally, the coordinates of the clusters of the eye fixations in Mono and Surround condition, differ at approximately 132 pixels (mean value). Additionally, we studied the positions of all clusters across all the frames in each video. Figure 2 and 3 show the positions of these clusters in screen coordinates. However, since these values do not provide any additional information about the evolution of the dispersion over time and whether the condition of the sound influences particular aspects of the clusters we performed an in depth analysis in each video.

2) Temporal analysis: Furthermore, we analyzed the evolution of the dispersion across the frames of all videos. Figure 4 shows how dispersion is evolved through time in the respective videos. In most of the videos the initial distance between the clusters is relatively small. Since before each video a fixation cross was presented to the subjects, approximately the first 10 to 20 frames showed no significant distance between the clusters in Mono and Surround condition. However, dispersion in Video 1 showed an odd pattern: after the first 30 frames, dispersion increases almost linearly reaching a local maximum before the first 100 frames ($\simeq$ 3 s). Subsequently, dispersion decreases and at approximately every 100 frames there is a local maximum again. This fluctuation is observed in other videos too but it is not that apparent as can be seen in Appendix B.

To investigate whether this behavior occurs because of the properties of the videos or the sound condition, we extracted the frames that correspond to these dispersion peaks. Figure 5 shows the six frames of Video 1 in which dispersion between the clusters of the Mono and Surround condition had the highest distance. We chose this video in particular since the periodicity observed is more evident compared to the other videos. As seen in Table 1 this video shows a street in which cars are passing by. It was observed that in the frames in which dispersion is high there is a relative high motion in the video (cars passing by). The same effect was also observed in Video 8 where there are also moving objects. In average,

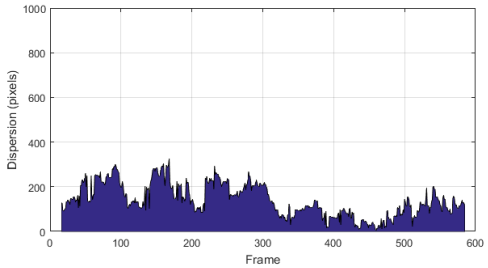


(a) Positions of clusters in Video 6 watched with Mono sound

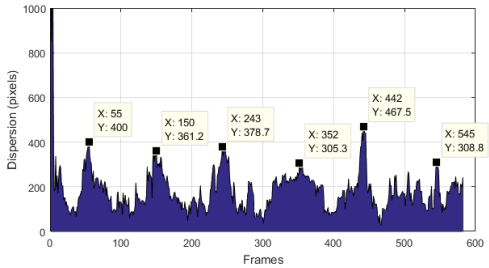


(b) Positions of clusters in Video 6 watched with Surround sound

Fig. 3: Clusters of all frames in Video 6 for Mono and Surround condition.



(a) Temporal evolution of the dispersion in Video 6



(b) Temporal evolution of the dispersion in Video 1

Fig. 4: Evolution of dispersion through time in Video 6 (a) and Video 1 (b). Points in Video 1 represent the local maxima.

after 25 frames where an object is appearing in the scene, the dispersion has its (local) highest value.

However, in the videos where there was a constant motion of the camera or multiple objects moving in the scene (ex. Video 10) the periodicity is not that clear. There are still frames where dispersion is peaking but it remains high throughout the duration or does not show a clear pattern.

In order to further investigate the relationship of dispersion and motion in the videos with distinctive moving objects, we performed a point biserial correlation test. We labeled each frame of the videos as 1 when there is a relative moving object in the scene and 0 when not. The aforementioned correlation analysis measures the strength of association between a continuous-level variable (dispersion values) and a binary (motion in frame) variable. The result was a positive correlation, $r = 0.084, n = 581, p = 0.043$. Overall, higher dispersion is associated with the higher levels of the group membership variables (frames with moving objects).

Nonetheless, there are moments where dispersion is relatively low. This means that the clusters between the Mono and Surround condition are close to each other. To further explore this behavior we extracted the frames that correspond to the local minima of the dispersion value in Video 1 (frames $36, 113, 194, 301, 466$ and 531). As can be seen in Appendix A these frames correspond to the time when a moving object starts appearing in the scene. This means that the sound condition did not influenced, or to a lesser degree, gaze allocation during that moments and will be further explained in the Discussion section.

B. Internal Dispersion

To further investigate the relationship between dispersion and sound condition we calculated the mean Intra Dispersion for each sound condition group for each video. We randomly split each sound condition group of 10 participants in two subgroups of 5 participants and calculated the Dispersion value per frame. We repeated this random split 10 times and took the mean Dispersion for each frame. The evolution of Intra Dispersion across time for each sound condition can be seen in Appendix B for all videos. Additionally we ran 2-tailed unequal variances paired samples t-test between 3 conditions:

· Intra Mono Dispersion - Intra Surround Dispersion
· Intra Mono Dispersion - Comparative Dispersion
· Intra Surround Dispersion - Comparative Dispersion

Table 5 presents the p value for each test and for each video. In general, there is a statistically significant difference between the Intra Mono and Intra Surround Dispersion values for all videos (for example in Video 2: $t = 5.315, p < 0.05$). However, in Video 1 and 4 there is no statistically significant difference between the Intra Surround and Comparative Dispersion values ($t = -4.043, p > 0.05$). Additionally, fixations in the Intra Surround condition are less dispersed ($M = 168$) compared to the Intra Mono condition ($M = 192$).

C. Questionnaire

In order to validate that our technique to provide the surround sound content was appropriate, we asked the participants to fill a questionnaire regarding the virtual azimuthal location of the sound sources playing. The subjects had to listen to one sound at a time playing from only one virtual speaker. If their response matched the actual location of the speaker it was

TABLE V: P value of paired t-tests in three conditions

Video ID	P value of two - tailed paired t-tests		
	Intra Mono - Intra Surround	Intra Mono - Comparative	Intra Surround - Comparative
1	0.000003	0.00006	0.105298
2	0	0	0
3	0	0	0.450241
4	0.005763	0	0
5	0.028099	0	0
6	0.000128	0.045359	0.002672
7	0.00284	0	0
8	0	0.020981	0
9	0	0	0
10	0	0	0

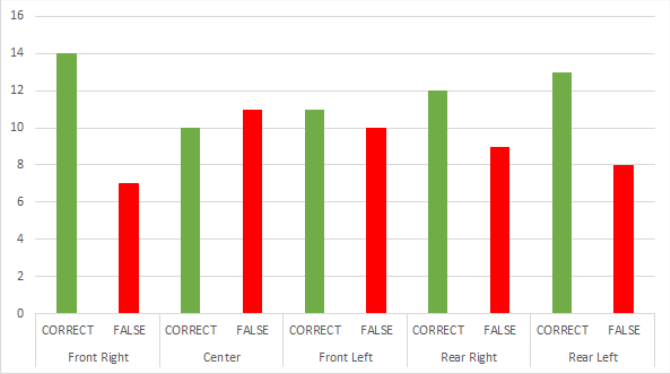


Fig. 5: Correct and false identifications of sound source ($n = 21$).

counted as correct. As seen in Fig. 5: center, front left and rear right virtual speakers were not perceived correctly. On average 10 out of the 21 participants perceived the virtual speakers in incorrect azimuthal locations. In general, front right and rear left speakers were the only ones that were perceived more accurately (14 and 13 out of 21 participants respectively).

D. Auditory stimuli and fixations

In order to investigate whether the location of the virtual sound source influenced the allocation of the fixations, we performed a one-way MANOVA between the pixel coordinates of the clusters in Surround condition as the dependent variables and the respective speakers playing at that frame as the independent variable for all videos. The results revealed that there is a statistically significant difference in the positions of the clusters based on the speaker playing $F(8, 156) = 34.446, p < 0.0005, Wilk's\Lambda = 0.652, partial\eta^2 = 0.192$. This means that the locations of the clusters are significantly dependent on which speaker is active.

Furthermore, a post hoc Tukey test showed that the pixel coordinates of the clusters were not statistically significantly different between the rear left and rear right speaker respectively ($p = 0.992$).

V. Discussion

In this study we compared the positions of the clusters of fixations of multiple participants while freely looking at videos with Mono and Surround sound respectively. We found that, in general, clusters in the Surround condition are positioned 132 pixels far from the ones in Mono. However, in moments where moving objects start appearing in a scene, the influence of sound, both in Mono and in Surround condition, is outweighted by visual information.

Our results suggest that these moments when attentional synchrony between the two clusters is high, sound condition has the same impact in the allocation of gaze. Although the clusters after these moments are dissociated with each other, we cannot deduce whether this behavior has occurred because of the sound condition solely. Our primary limitation caused by the refresh rate of the eye tracking device (60 Hz) did not allow further investigation in the positions and amplitude of the saccades. The experiments performed by Coutrot et al. [18] showed that "sound strengthens visual salience". That hypothesis was confirmed by the results of their saccade amplitude distribution analysis. As we could not perform such experiment we can only assume that surround sound has a general impact in the positions of the fixations when there is no motion contrast in the scene.

Furthermore, as assumed by Mital et al. "the increase in visual feature contributions during attentional synchrony may suggest that gaze is involuntary captured by sudden unexpected visual features such as object appearances or motion onsets" [7]. Our results conform with this suggestion as especially in these moments sound condition did not influence the allocation of the fixations and clusters were positioned in the screen area where the object appeared. However, when there is no abrupt onset of a new object in a scene, sound condition modulates gaze allocation. This can be seen from the results of the Internal Dispersion analysis that showed that Intra Dispersion in Mono condition is significantly different from the Intra Dispersion in Surround. This suggests that either the eye movements in both sound conditions are random, which is not likely to happen because of the strength of the visual stimuli features, or that the sound condition indeed modulates fixations allocation.

The dependency found between the positions of the clusters and the active speaker in the Surround condition strengthens the above assumption. Nonetheless, as our spatialization technique did not properly provide the surround sound content (as seen in the responses of the questionnaire) we cannot assume that this dependence yields that gaze is steered towards the active speaker, as it has been shown in experiments performed by Quigley et al. [25]. Additionally, this dependence is only apparent for the front panel of the virtual speakers (front right, front left and center) as the pixel coordinates of the clusters in rear left and rear right speakers were not significantly different.

Some further aspects of the auditory stimuli also warrant discussion. The sounds used in the experiments were deliberately non related to the visual stimuli. As assumed "a readily identifiable sound might provide a more complex spatial cue by virtue of the listener's world knowledge" [25]. We believe that the choice of using dissociated sounds to the content of the videos provided non biased results. As object semantics are the main contributors to gaze allocation, sounds related to these objects would have created altered results. However, additional properties of these sounds, as amplitude, frequency

and duration can be further explored in a surround system context during dynamic scene viewing.

The aforementioned limitations regarding the sample rate of the eye tracking device along with the malfunction of the spatialization technique can provide substantial experience on the future development of such experiments.

First and foremost, the use of a physical surround setup in the experimental procedure seems necessary. While other studies have managed to provide accurate surround content through headphones (for example Quigley et al. [25]), our technique failed to address the issue. As we can not be sure whether it was the engine used to create the spatialization or the headphones used to deliver the surround content - or even both - that did not function properly, we can only hypothesize that by eliminating all these factors the problem would have eradicated.

Additionally, as far as the methodology is concerned, we believe that another group of participants should have been added in the experimental procedure, where only the sound condition would have been tested in a black screen (screen with no visual features at all). Previous studies have used this methodology to evaluate the evolution of dispersion between auditory (A), visual (V) and unimodal (AV) conditions (for example, Coutrot et al. [18]; Quigley et al. [25]; Song et at. [26]; and Song et al. [20]). As a result, we would have a better baseline to compare the dispersion values of the clusters from the videos in both sound conditions.

VI. Conclusion

Our study constitutes a first attempt to understand if surround sound has an impact on gaze allocation during video exploration. While surround sound has been researched within the static images domain, dynamic scenes have not been explored. We showed that with Surround sound, the eye fixations are broadly far from the ones with Monophonic sound. Moreover, gaze allocation is dependent on the active speaker in each frame with the exception of rear left and rear right speakers. Our results highlighted that the effect is not apparent when abrupt moving objects are presented into the scene. All these results indicate that further investigation of spatial auditory and visual features in dynamic scenes should be further explored. A better experimental design that resolves the issues regarding the eye tracking device and the spatialization technique will provide a more accurate answer on how surround sound impacts gaze distribution.

References

[1] M. Carrasco, "Visual attention: The past 25 years," *Vision Research*, vol. 51, pp. 1484–1525, 2011.

[2] D. J. Parkhurst and E. Niebur, "Scene content selected by active vision," *Spatial Vision*, vol. 16, no. 2, pp. 125–154, 2003.

[3] P. Lennie, "The cost of cortical computation," *Current Biology*, vol. 13, no. 6, pp. 493–497, 2003.

[4] G. Rizzolatti, L. Riggio, I. Dascola, and C. Umilta, "Reorienting attention across the hotizontal and vertical meridians: evidence in favor of a premotor theory of attention," *Neuropsychologia*, vol. 25, no. 1, pp. 31–40, 1987.

[5] A. V. Belopolsky and J. Theeuwes, "When are attention and saccade preparation dissociated?" *Phycological Science*, vol. 20, no. 11, pp. 1340–1347, 2009.

[6] R. M. Klein and A. Ponterfact, "Does oculomotor readiness mediate cognitive control of visual attention?" in *Attention and Performance XV: Conscious and Nonconscious Information Processing*, C. Umilta and M. Moscovitch, Eds. The MIT Press, 1980, ch. 13, pp. 333–350.

[7] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cognitive Computation*, vol. 3, no. 1, pp. 5 –24, 2011.

[8] J. M. Henderson, "Regarding Scenes," *Current Directions in Psychological Science*, vol. 16, no. 4, pp. 219–222, 2007.

[9] J. R. Anderson, "Attention and performance," in *Cognitive psychology and its implications*, 8th ed. New York: Worth Publishers, 1990, pp. 53–77.

[10] S. Mannan, D. Wooding, and K. Ruddock, "The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images," *Spatial Vision*, vol. 10, no. 3, pp. 165–188, 1996.

[11] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, no. 5, pp. 643–659, 2005.

[12] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," *Vision Research*, vol. 46, no. 26, pp. 4333–4345, 2006.

[13] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[14] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[15] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence*, L. Vaina, Ed. Dordrecht: Springer Netherlands, 1987, ch. 4, pp. 115–141.

[16] H. E. Egeth and S. Yantis, "Visual attention: control, representation, and time course," *Annual Review of Psychology*, vol. 48, no. 1, pp. 269–297, 1997.

[17] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2843–2498, 2007.

[18] A. Coutrot, N. Guyader, G. Ionescu, and A. Caplier, "Influence of soundtrack on eye movements during video exploration," *Journal of Eye Movement Research*, vol. 5, no. 4, pp. 1–10, 2012.

[19] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes." *Journal of Vision*, vol. 14, no. 8, pp. 1–17, 2014.

[20] G. Song, D. Pellerin, and L. Granjon, "Different types of sounds influence gaze differently in videos," *Journal of Eye Movement Research*, vol. 6, no. 4, pp. 1–13, 2013.

[21] H. Zou, H. J. Muller, and Z. Shi, "Non-spatial sounds regulate eye movements and enhance visual search," *Journal of Vision*, vol. 12, no. 5, pp. 1–18, 2012.

[22] P. A. Arndt and H. Colonius, "Two stages in crossmodal saccadic integration: evidence from a visual-auditory focused attention task," *Experimental Brain Research*, vol. 150, no. 4, pp. 417–426.

[23] B. D. Corneil and D. P. Munoz, "The influence of auditory and visual distractors on human orienting gaze shifts." *The Journal of Neuroscience*, vol. 16, no. 24, pp. 8193–8207, 1996.

[24] B. D. Corneil, M. Van Wanrooij, D. P. Munoz, and A. J. Van Opstal, "Auditory-visual interactions subserving goal-directed saccades in a complex scene." *Journal of Neurophysiology*, vol. 88, no. 1, pp. 438–54, 2002.

[25] C. Quigley, S. Harding, M. Cooke, P. König, and S. Onat, "Audio-visual integration during overt visual attention," *Journal of Eye Movement Research*, vol. 14, no. 2, pp. 1–17, 2008.

[26] G. Song, D. Pellerin, and L. Granjon, "Sound effect on visual gaze when looking at videos," in *19th European Signal Processing Conference*, 2011, pp. 2034–2038.

[27] M. Boltz, "The cognitive processing of film and musical soundtracks," *Memory & Cognition*, vol. 32, no. 7, pp. 1194–1205, 2004.

[28] M. Bindemann, "Scene and screen center bias early eye movements in scene viewing," *Vision Research*, vol. 50, no. 23, pp. 2577–2587, 2010.

[29] P. H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of Vision*, vol. 9, no. 7, pp. 1–16, 2009.

[30] Y. Sawahata, R. Khosla, K. Komine, N. Hiruma, T. Itou, S. Watanabe, Y. Suzuki, Y. Hara, and N. Issiki, "Determining comprehension and quality of TV programs using eye-gaze tracking," *Pattern Recognition*, vol. 41, no. 5, pp. 1610–1626, 2008.

APPENDIX A

(a) Frame 36

(b) Frame 113

(c) Frame 194

(d) Frame 301

(e) Frame 466

(f) Frame 531

Fig. A.1: Frames that correspond to the local minima of Video 1

APPENDIX B

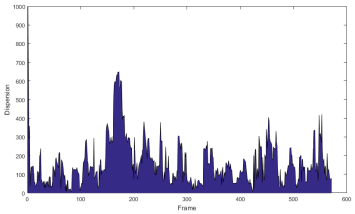


Fig. B.2: Video 1 Intra Mono Dispersion

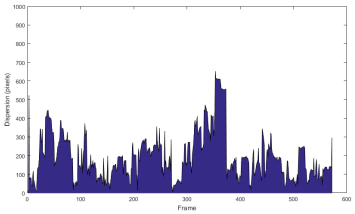


Fig. B.3: Video 1 Intra Surround Dispersion

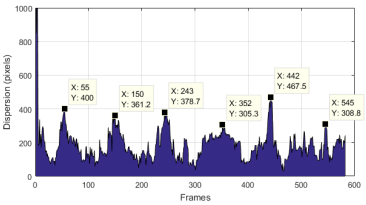


Fig. B.4: Video 1 Comparative Dispersion

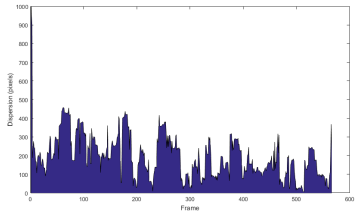


Fig. B.5: Video 2 Intra Mono Dispersion

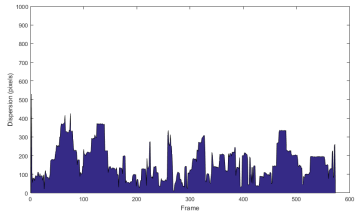


Fig. B.6: Video 2 Intra Surround Dispersion

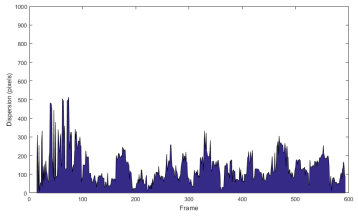


Fig. B.7: Video 2 Comparative Dispersion

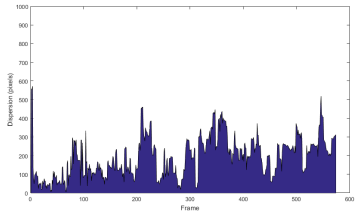


Fig. B.8: Video 3 Intra Mono Dispersion

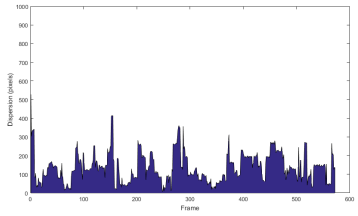


Fig. B.9: Video 3 Intra Surround Dispersion

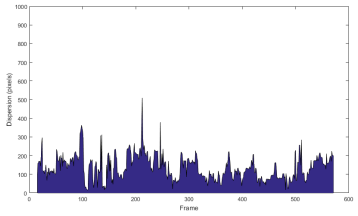


Fig. B.10: Video 3 Comparative Dispersion

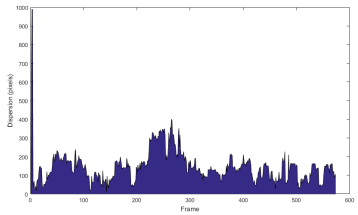


Fig. B.11: Video 4 Intra Mono Dispersion

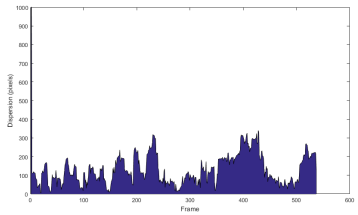


Fig. B.12: Video 4 Intra Surround Dispersion

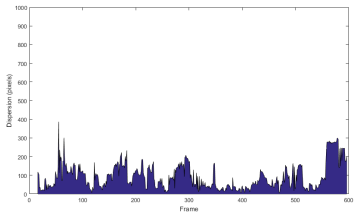


Fig. B.13: Video 4 Comparative Dispersion

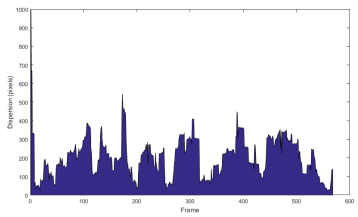


Fig. B.14: Video 5 Intra Mono Dispersion

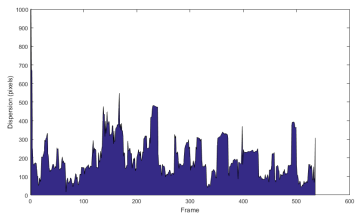


Fig. B.15: Video 5 Intra Surround Dispersion

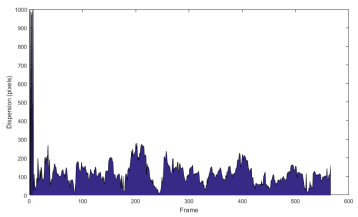


Fig. B.16: Video 5 Comparative Dispersion

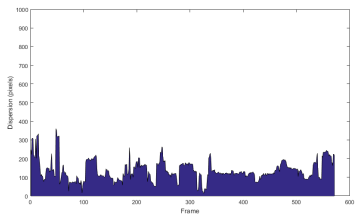


Fig. B.17: Video 6 Intra Mono Dispersion

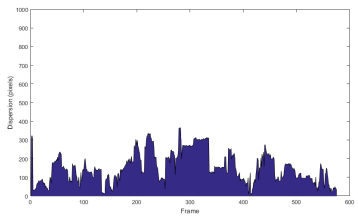


Fig. B.18: Video 6 Intra Surround Dispersion

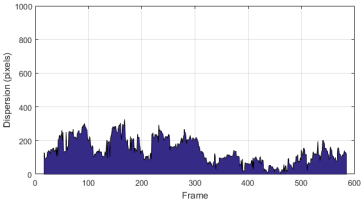


Fig. B.19: Video 6 Comparative Dispersion

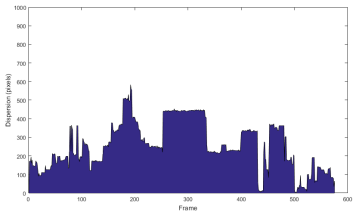


Fig. B.20: Video 7 Intra Mono Dispersion

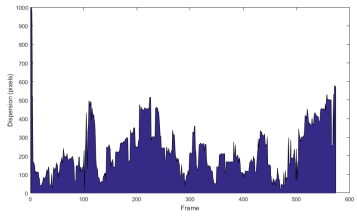


Fig. B.21: Video 7 Intra Surround Dispersion

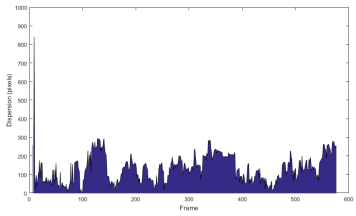


Fig. B.22: Video 7 Comparative Dispersion

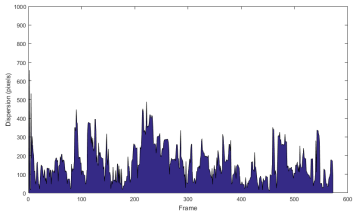


Fig. B.23: Video 8 Intra Mono Dispersion

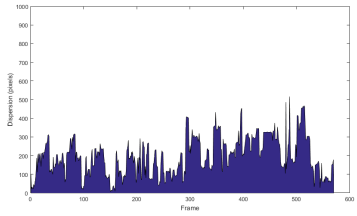


Fig. B.24: Video 8 Intra Surround Dispersion

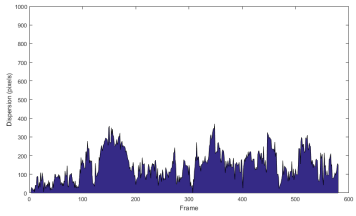


Fig. B.25: Video 8 Comparative Dispersion

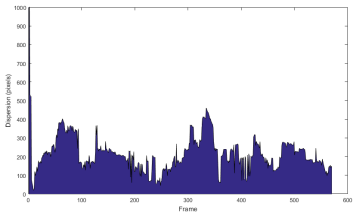


Fig. B.26: Video 9 Intra Mono Dispersion

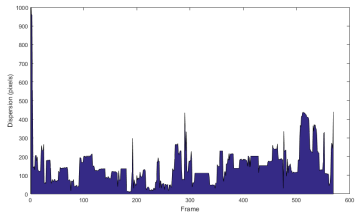


Fig. B.27: Video 9 Intra Surround Dispersion

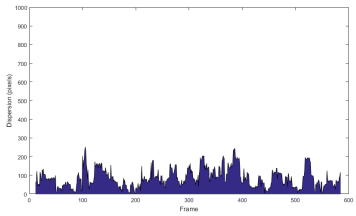


Fig. B.28: Video 9 Comparative Dispersion

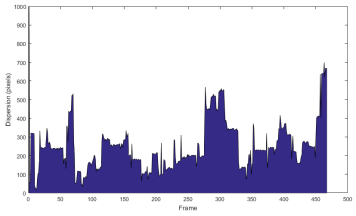


Fig. B.29: Video 10 Intra Mono Dispersion

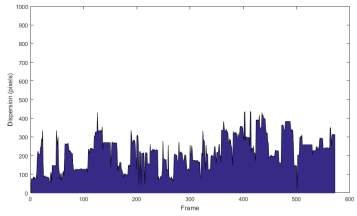


Fig. B.30: Video 10 Intra Surround Dispersion

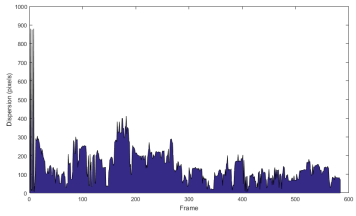


Fig. B.31: Video 10 Comparative Dispersion