Eyes ’n Ears:
Face Detection Utilizing Audio and Video Cues

B. Kapralos^{1,3}*, M. Jenkin*^{1,3}*, E. Milios*^{2,3} *and J. K. Tsotsos*^{1,3}

[1] Dept. of Computer Science, York University, Toronto, Ontario, Canada. M3J 1P3
[2] Dept. of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada. B3H 1W5
[3] Centre for Vision Research, York University, Toronto, Ontario, Canada. M3J 1P3

{billk, jenkin, eem, tsotsos}@cs.yorku.ca

Abstract

This work investigates the development of a robust and portable teleconferencing system utilizing both audio and video cues. An omni-directional video sensor is used to provide a view of the entire visual hemisphere thereby providing multiple dynamic views of the participants. Regions of skin are detected using simple statistical methods, along with histogram color models for both skin and non-skin color classes. Skin regions belonging to the same person are grouped together. Using simple geometrical properties, the location of each person’s face in the “real world” is estimated and provided to the audio system as a possible sound source direction. Beamforming and sound detection techniques with a small, compact microphone array allows the audio system to detect and attend to the speech of each participant, thereby reducing unwanted noise and sounds emanating from other locations. The results of experiments conducted in normal, reverberant environments indicate the effectiveness of both the audio and video systems.

1 Introduction

With the advent of the “Global Village”, teleconferencing has found a wide range of applications. From facilitating business meetings to aiding in remote medical diagnoses, it is used by corporate, university, medical, government and military organizations. Teleconferencing enables new operational efficiencies resulting in reduced travel costs, faster business decision making, increased productivity, reduced time to market and remote classroom teaching [3]. Various commercial teleconferencing systems exist, including basic static systems for use by two participants (one at each end of the connection). There are also systems intended for multiple speakers (see [9]), however, these systems provide a limited number of static or manually tracked views of the participants. As a consequence, in a multiple speaker setting, either a speaker must move into the camera’s view or a camera operator must manually track the speaker. This is both bothersome and inconvenient for the participants and has deterred many from using such systems. Furthermore, the presence of a camera operator in a teleconferencing session may perturb the group dynamics [12]. Vision based systems capable of detecting and tracking humans could be used to automate this task, however, such systems typically employ normal camera lenses, which capture only a narrow field of view. Choosing “where to look next” when a potential speaker is outside of the camera’s narrow field of view is a particularly complex task. Teleconferencing systems must be able to capture and transfer audio and video. As a result, in a multiple speaker setting, the teleconferencing system must be able to localize a speaker in the audio domain as well. However, it is hard to localize speakers based on audio cues alone. Background noise, other speakers and multi-pass reflections make sound source localization an extremely difficult task. Although sound localization systems exist, most rely on extensive microphone arrays [1, 4, 11] which require expensive specialized equipment and are computationally intensive.

Rather than attempting to solve these complex problems independently, our research investigates the development of a teleconferencing system which integrates both audio and visual cues. By taking advan-

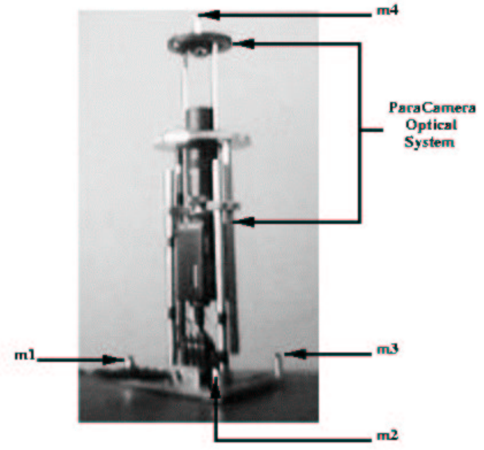


Figure 1: Combined Eyes 'n Ears Audio and Video Sensor

tage of the unique features and properties of the audio and video domains, we hope to overcome the inherent disadvantages of each, resulting in a combined sensor system which is more effective than either sensor is individually. The ultimate goal of the Eyes 'n Ears project is to develop an affordable, low maintenance and portable video teleconferencing system capable of locating and tracking a speaker in a multiple speaker setting.

1.1 Overview

Figure 1 illustrates the combined audio and video sensors comprising the Eyes 'n Ears hardware. The system is compact, lightweight and portable and is meant to be placed in the middle of a table with the participants of the teleconference session seated around it. The video system determines potential "real world" positions of (or directions to) each speaker and then provides this information to the audio system. Given this information, using beamforming [6] and sound detection techniques, the audio system detects and focuses on the speech of each potential speaker, rejecting any false positives identified by the video system. The following sections describe the operation of the audio and video systems in greater detail.

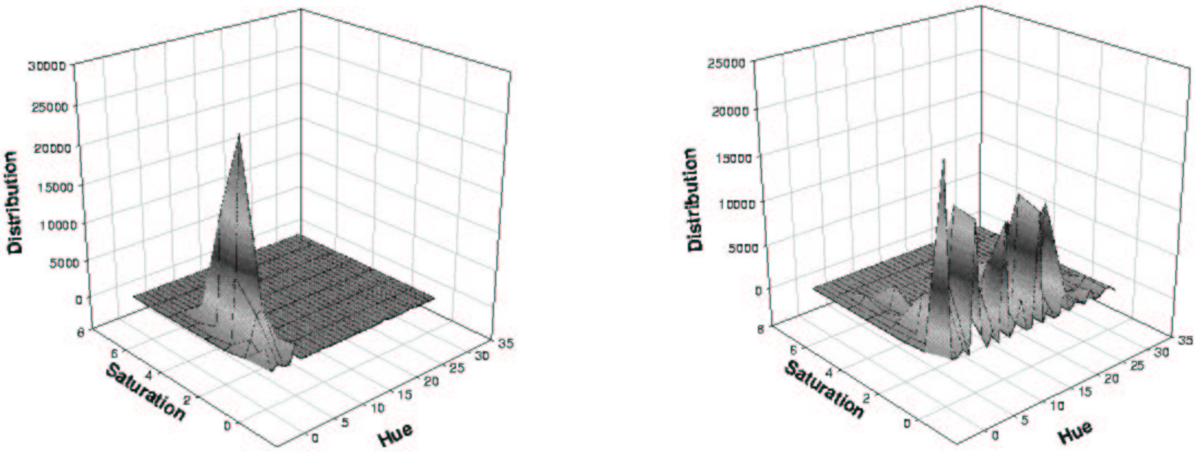
2 Video System: Skin Pixel Classification

Cyclovision's ParaCamera omni-directional optical system [8] is utilized. The ParaCamera consists of a high precision paraboloidal mirror and a combination of special purpose lenses. By aiming a camera to the face of the paraboloidal mirror, the combination of these optics permit the ParaCamera to capture a 360^o view of potential speakers from a single viewpoint. Two-dimensional Hue-Saturation histograms for both skin and non-skin color classes were constructed by manually classifying portions of images obtained with the ParaCamera, as either skin or non-skin. Each histogram contains a total of 256 *bins* (32 values for hue and eight for saturation). 88,888 skin pixels obtained from the portions of exposed skin regions from 30 subjects of various ethnic groups were used to obtain the skin class. 223,728 non-skin pixel samples obtained from image regions which did not contain any exposed human skin were used to construct the non-skin class. Figure 2 illustrates the Hue-Saturation distribution of the skin and non-skin classes. Given the skin and non-skin histograms, following [7], Bayesian probability can be used to classify pixels within each incoming ParaCamera image as either being skin or non-skin.

Once pixels are classified as skin or non-skin, erosion and dilation operators are applied to remove isolated pixel regions. The remaining pixels are then grouped into labeled regions using an 8-neighbor connected components operator and any components smaller than a pre-defined threshold size are eliminated. A search is then conducted to cluster connected regions which are spatially close. Assuming there is a reasonable amount of space between participants in view, each cluster of skin regions is assumed to correspond to a particular person. Given the geometry of the ParaCamera, the region of each cluster furthest from the center of the ParaCamera image is chosen as the face. Once each face has been found, an estimate of its position (or direction) in the real world is made and provided to the audio system.

2.1 Converting to World Coordinates

To determine locations from a single ParaCamera image to locations in the real world, a ground-plane perpendicular to the optical axis of the ParaCamera is assumed [2]. Informal lab surveys suggest the average height of seated people is $1.20m$ above the floor. As a result, the ground-plane ("head-plane") for this application is chosen to be $1.20m$ above the floor. Real world locations are determined using a method similar to that described in [5]. Essentially, the location is the point of intersection, obtained by extending the line between the focus of the ParaCamera's paraboloidal



(a): Skin Histogram

(b): Non-Skin Histogram

Figure 2: Hue-Saturation Histograms for Skin and Non-Skin Color Classes.

mirror and the reflection point of the participants face on the mirror until it intersects with the ground-plane. The ground-plane assumption can be relaxed when a direction to the participant, as opposed to a location, is of interest. This corresponds to the *far* field acoustical model described in the following section.

3 Audio System: Beamforming

Beamforming takes advantage of the time delay between the arrival of sound to each of the microphones in a microphone array. The Eyes 'n Ears sensor utilizes four microphones $(m_1, \dots, m_4)$, with m_1 chosen as the array reference point (origin). Applying an appropriate time delay Δ_i to the signal received by the remaining three microphones $(m_2, \dots, m_4)$, directs the microphone array to particular direction and distance. This tunes the array to a particular sound source while attenuating noise or signals propagating from other directions and locations.

Given the position of (or direction to) a potential sound source obtained by the vision system, the audio system is tuned to the particular location (direction). For beamforming, there are different solutions depending on whether the sound source is in the *near* or *far* field. Beamforming when the sound source is in the far field is simpler, computationally "less expensive" and does not require the exact sound source location but rather only the direction of propagation. In contrast, beamforming for a sound source in the near field is computationally more complex and requires the location of the sound source. Generally, in a teleconferencing session the participants will be located close to the sensor (e.g. in the near field). However, in order to exploit the advantages associated with a far field sound source, rather than assuming a near field source, the error in assuming a far field source is calculated using the technique described in [6]. When this error is below a pre-defined threshold value, the sound source is considered to be in the far field, otherwise a near field source is assumed.

Far Field Source: Given the source's direction of propagation β and the position of each microphone relative to the array origin x_i $(i = 2, \dots, 4)$, simple geometry may be used to directly determine the time difference (Δ_i) for each of the three microphones relative to the array origin

$$\Delta_i = -\frac{\beta \cdot x_i}{v_{sound}}$$

where β is the unit vector denoting the direction of propagation relative to the array's origin, and the speed of sound v_{sound} is assumed to be constant at $345 m/s$.

Near Field Source: With a near field source, the delay required for the signal of microphone m_i is related to the difference in distance between the sound source and the array reference $d_{s,1}$ and the sound source and the i^{th} microphone $(d_{s,i})$

$$\Delta_i = \frac{d_{s,1} - d_{s,i}}{v_{sound}}$$

The delay may also be determined in terms of the sound source location x_s

$$\Delta_i = \frac{||x_s||^2}{v_{sound}}\left[1-\sqrt{1+\frac{||x_i||^2}{||x_s||^2}-2\frac{x_s\cdot x_i}{||x_s||^2}}\right]$$

3.1 Sound Detection

Since speech is the signal of interest, the signal received by each of the microphones is filtered using a FIR filter to attenuate signals falling outside of the 200–4000Hz frequency regions associated with human speech [10].

The beamformed signal s_{beam} is obtained by summing the appropriately delayed signals of each microphone

$$s_{beam} = \sum_j \sum_{k=1}^{k=4} s_k[j+\Delta_k]$$

where $\Delta_1 = 0$. Signal magnitude E_{signal} is computed as

$$E_{signal} = \frac{\sum_{j=0}^{N} |x[j]|}{N}$$

and may be used as an indication of the presence or absence of a sound source in the environment (e.g. the magnitude of a signal associated with a sound source will generally be greater than the magnitude of a signal obtained in the absence of a sound source). However, signal variance provides a stronger and more reliable indication to the presence or absence of a sound source. Rather than computing the variance of an entire signal window (2048 samples), variance is computed by dividing the window into $M = 56$ sub-windows of $s = 32$ samples each. The variance of each sub-window is computed and finally, the variance for the entire window is computed by taking the mean of the 52 sub-window variances

$$V_{signal} = \frac{\sum_{j=0}^{j=M-1}\sum_{k=j\times s}^{k=j+s-1}\frac{|x[k]-\bar{x}|}{s-1}}{M}$$

where the values for M and s were chosen through informal testing. This averaging process leads to a reduction of noise present in the original sample window [6], providing a strong indication to the presence or absence of a sound source.

Rather than relying on the absolute magnitude of the beamformed signal (which may vary without a noticeable change in the background noise or sound source level), as a measure of the source location as done in many beamforming applications, the beamformed signal s_{beam} is compared to the average signal of the microphones s_{avg}

$$s_{avg} = \sum_j (s_1[j]+s_2[j]+s_3[j]+s_4[j])$$

and signal difference s_{dif}, is computed as follows

$$s_{dif} = \begin{cases} \frac{s_{beam}-s_{avg}}{E_{mean}} & \text{if } (s_{beam}-s_{avg}) > 0 \\ 0 & \text{if } (s_{beam}-s_{avg}) \leq 0 \end{cases}$$

where E_{mean} is the mean of 20 consecutive average signal magnitudes (E_{avg}) values obtained after every 500 iterations. s_{dif}, the normalized difference between s_{avg} and s_{beam}, is maximized when the beamformer is focused at the location corresponding to the sound source. Provided the value of s_{dif} and V_{beam} corresponding to the location (direction) of a face are above some pre-defined threshold values Dif_{thresh} and V_{thresh} respectively, the presence of a speaker can be confirmed.

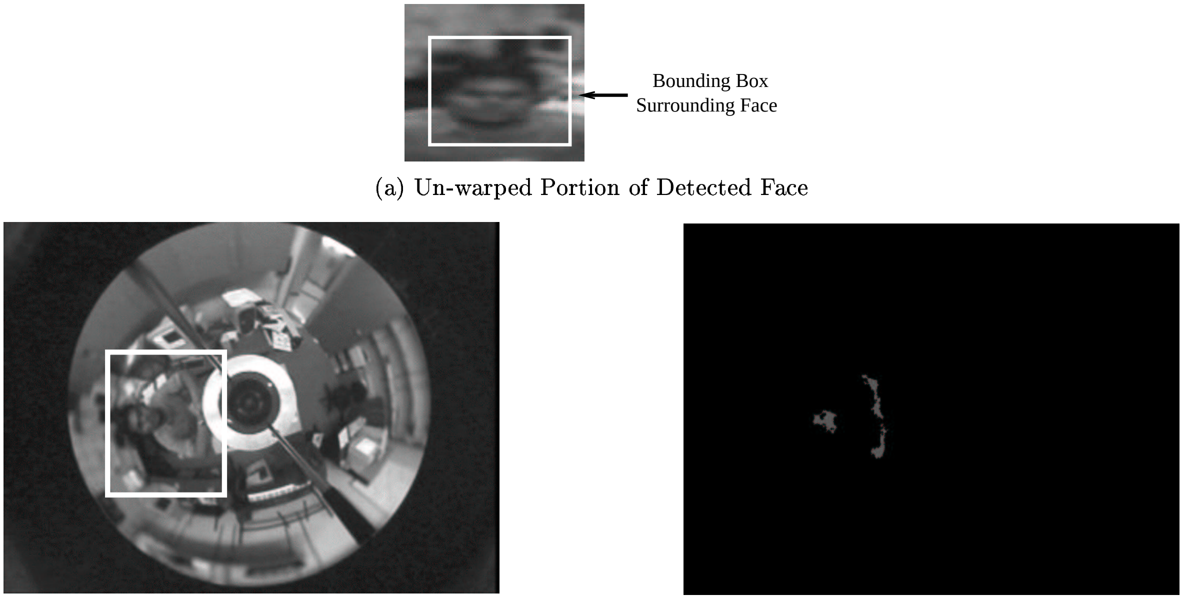
4 Results

4.1 Video System Example

A sample of the face detection process is provided in the following sequence of images shown in Figure 3. Figure 3c illustrates the result of applying the skin pixel classification process to the image of Figure 3b. As shown in Figure 3a, the face present in the Para-Camera image of Figure 3b is correctly detected.

4.2 Audio System

Two loudspeakers continuously outputting a recording of a male subject reading a phrase were placed in a room. The first sound source (sound source A) was placed at $x = 0.80$, $y = 0.30$ and $z = 0.10$ relative to m_1 while the second sound source (sound source B) was placed at $x = -0.40$, $y = 0.80$ and $z = 0.12$ relative to m_1. A region of the room ($2.8m \times 2.80m$ at a fixed height, z) was divided into voxels, where each voxel $0.10m \times 0.10m \times 0.01m$. The beamformer was focused to the center of each voxel and the appropriate measurements obtained. This was repeated for several values of z, ranging from $0.00m$ to $0.30m$. The



(a) Un-warped Portion of Detected Face

(b) ParaCamera Image

(c) Skin Classification

Figure 3: Example of the Face Detection Process

Value	**Mean**	**Stnd. Dev**	Max.	**Min.**
E_{beam}	23.59	11.77	81.23	2.99
E_{avg}	26.27	12.379	73.35	2.97
V_{beam}	15.05	12.34	83.43	0.87
V_{avg}	37.20	1.28	73.35	0.91
s_{dif}	-0.20	0.23	1.03	-1.28

Table 1: Audio System Experiment Summary. The Presence of Two Sound Sources

experimental results are summarized in Table 1. The mean value for s_{dif} is 0.10 with a standard deviation of 0.23.

Of the 23,520 distinct locations which the sensor was focused to, Table 2, lists the 11 locations with corresponding values of $s_{dif} > Dif_{thresh}$. All locations (except one), are close to the location of sound source A (within 0.40m, 0.30m and 0.20m on the x, y and z axis respectively). The value of s_{dif} at the actual location of sound source A is 0.93, greater than $Dif_{thresh} = 0.90$. Although the value of s_{dif} corresponding to the location of sound source B is less than Dif_{thresh}, $s_{dif} > Dif_{thresh}$ for the location $x = -0.40m$, $y = 0.90m$ and $z = 0.24m$, very close to the location of sound source B, leading to an error of 0.00m, 0.10m and 0.12m for the x, y and z axis respectively. In addition, the value of V_{avg} is also greater than the $V_{thresh} = 6.00$ indicating the presence of a sound source.

s_{dif}	V_{avg}	**x**	**y**	**z**
1.03	8.96	0.8m	-0.6m	0.13m
0.98	9.54	0.7m	-0.6m	0.13m
0.96	55.20	-0.4m	0.5m	0.24m
0.95	25.25	1.2m	-0.5m	0.08m
0.94	27.16	0.5m	-0.3m	0.24m
0.93	29.60	0.8m	-0.4m	0.11m
0.93	57.51	1.2m	-0.7m	0.27m
0.92	31.60	0.8m	-0.5m	0.19m
0.91	63.66	1.2m	-0.6m	0.06m
0.90	57.86	0.7m	-0.5m	0.07m
0.90	55.86	0.8m	-0.6m	0.07m

Table 2: Locations In Experiment Two Corresponding to $s_{dif} > Dif_{thresh}$

Figures 4a and 4b illustrate the values of s_{dif} before applying the threshold operation, at $z = 0.13$ and $z = 0.24$ respectively, while Figures 5a and 5b illustrate the result of applying the threshold operation to the values shown in Figures 4a,b.

The results of the previous experiment indicate the audio system is capable of detecting the presence of a sound source as well as locating its position (within an error bound). The following experiment describes the behavior of the audio system in the absence of any sound sources. The procedure performed in the previous experiment was repeated in the absence of a sound source. A summary of the experimental results are shown in Table 3. The mean s_{dif} is 0.02 with a standard deviation of 0.03 and maximum value is 0.22, well below the threshold value of $Dif_{thresh} = 0.90$. Similarly, the maximum variance V_{avg} value encountered is 1.08, once again well below the average variance threshold of $V_{thresh} = 6.00$. As a result, the value of V_{avg} measured at each location is less than V_{thresh}, allowing the audio system to easily detect the absence of a sound source.

Value	**Mean**	**Stnd. Dev**	Max.	**Min.**
E_{beam}	4.09	1.97	17.44	0.96
E_{avg}	4.01	1.95	16.97	0.92
V_{beam}	0.81	0.08	1.37	0.64
V_{avg}	0.76	0.05	1.08	0.63
V_{avg}	0.01	0.03	0.22	-0.05

Table 3: Audio System Experiment Summary. The Absence of any Sound Source.

5 Discussion

This paper describes the audio and video components, developed for use in a multi-speaker teleconferencing session. Preliminary results indicate both systems are capable of performing their intended tasks accurately. Various factors may affect each component, however, these factors are usually specific to either the audio or video system. For example, a reverberant environment may result in the incorrect localization of a sound source, but will not affect the video system. Similarly, the color of objects in the environment has no bearing on the audio system whereas it may negatively affect the video system and lead to the incorrect classification of non-skin regions as skin. Finally, by locating the location (direction) of potential faces within the ParaCamera's view, the video system essentially reduces the "workspace" of the audio system from many directions to only a few (e.g. 1–10), making the audio system's task tractable. Although each system has its share of potential problems, combining audio and video cues allow many of the shortcomings inherent with each component to be improved or overcome, leading to more accurate and robust target detection. More extensive testing is currently being conducted to further evaluate the effectiveness of the system.

Acknowledgments

This research was generously supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), Communications and Information Technology Ontario (CITO) and Centre for Research in Earth and Space Technology (CRESTech). The authors would also like to thank Dan Zikovitz for his valuable feedback and comments.

References

[1] M. Brandstein, M. Adcock, and H. Silverman. A practical time-delay estimator for localizing sound sources with a microphone array. *Comput. Speech. Lang.*, 9:153–169, 1995.

[2] K. Danilidis. Personal communication.

[3] A. W. Davis. *Integrated Collaboration: Driving Business Efficiency into the Next Millennium.* Forward Concepts, 1999.

[4] J. Flanagan, D. Johnston, R. Zhan, and G. Elko. Computer steered microphone arrays for sound transduction in large rooms. *J. Acoust. Soc. Am.*, 78(2):1508–1518, 1985.

[5] D. Gutchess, A. Jain, and S. Cheng. Automatic surveillance using omni-directional and active cameras. In *Proc. Asian Conf. Comput. Vis.*, 2000.

[6] D. Johnson and D. Dudgeon. *Array Signal Processing: Concepts and Techniques.* Prentice Hall, USA, 1993.

[7] M. J. Jones and J. M. Rehg. Statistical color models with applications to skin detection. Technical Report CRL 98/11, Compaq Computer Corp., Cambridge, MA USA, 1998.

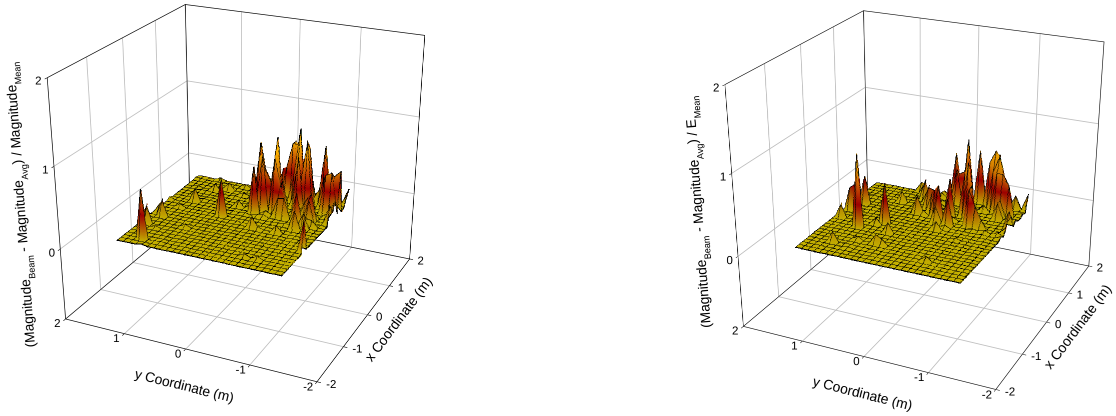
[8] S. Nayar. Omnidirectional video camera. In *Proc. DARPA Image Understanding Workshop*, pages 235–241, New Orleans, LA, 1997.

[9] Panasonic. Panasonic Corporation KXC-M7800 vision pro series 7800 video teleconferencing system.

[10] L. R Rabiner and M. R Sambur. An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*, 54(2), 1975.

[11] D. Rabinkin. Digital hardware and control for a beam-forming microphone array. Master's thesis, Department of Electrical Engineering, Rutgers University, New Brunswick NJ USA, January 1994.

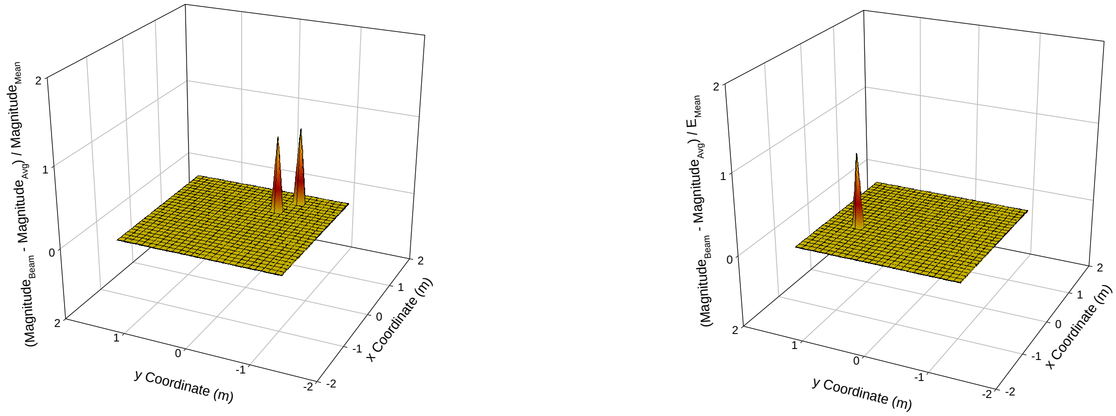
[12] R. Yong, A. Gupta, and J. Cadiz. Viewing meetings captured by an omni-directional camera. In *ACM Trans. Comput.-Hum. Interact.*, March 2001.



(a) No Threshold Applied, z = 0.13

(b) No Threshold Applied, z = 0.24

Figure 4: Signal Difference (s_{dif}) Before Applying Threshold Operation for z = 0.13 and z = 0.24.



(a) Threshold Applied, z = 0.13

(b) Threshold Applied, z = 0.24

Figure 5: Signal Difference (s_{dif}) After Applying Threshold Operation for z = 0.13 and z = 0.24.