DIRECTING ATTENTION IN ONLINE AGGREGATE SENSOR STREAMS VIA AUDITORY BLIND VALUE ASSIGNMENT

Robert Malkin, Datong Chen, Jie Yang, Alex Waibel

interACT @ Carnegie Mellon University
{rgmalkin,datong,yang+,ahw}@cs.cmu.edu

ABSTRACT

Multiparty collaborative applications in which groups of people act in concert to achieve some real-world goal abound. In these situations, it is useful for a central planning agent to receive online audiovisual information from all participants. However, as the size of the group grows, it becomes difficult to process all the sensory streams; cognitive overload prevents direct analysis of sensory streams for situational awareness. To avoid this situation, an automatic method is needed to assign value to each stream and direct the attention of the planning agent to those streams which are most valuable. We present an audio-based blind value assignment (BVA) method to address this problem, and experiments demonstrating the method's efficacy. We demonstrate that using audio BVA techniques results in automatic value judgments which are broadly similar to human value judgments and superior to automatic judgments based on video information.

1. INTRODUCTION

Multiparty collaborative applications in which groups of people act in concert to achieve some real-world goal abound. Firefighting, rescue, emergency response, and multiplayer games are examples. In these scenarios, there is often a planning agent whose role is to collect information from the group, construct a situation model, and adjust the plan accordingly. It is desirable in these scenarios to have not just verbal situation reports, but sensory information as well. Sensory information can be transferred by audiovisual capture devices worn by each group member; we refer to this set of streams as an online aggregate sensor stream (OAS). By transmitting this data, group members can be relieved of some of the burden of making verbal reports, and the planning agent can see and hear important parts of the situation firsthand. However, as group size grows, the amount of data being transferred becomes overwhelming and thus useless. To overcome this problem, the planning agent's attention should be focused on those data streams which are of interest. The problem of automatically determining the identities of these streams from a pool of candidates is called value assignment.

In this research, we consider the case of *blind* value assignment (BVA), in which we have no prior knowledge about the contents of the data and must rely on the statistics of the streams to judge value. Blindness means that we have no way of biasing the value assignment procedure toward classes of events which are known to be important for some application. However, blindness does provide robustness to noisy or unseen conditions under which event models may break down. BVA can thus always serve as a fast baseline value assignment method. Our approach to BVA focuses on audio information. We make this choice because audio has many attractive features for the task. Among these are low bandwidth and processing costs, omnidirectionality, and immunity to occlusion, sensor motion, and changes in lighting conditions. Furthermore, events which are important in the real world often leave behind clear acoustic evidence.

Our audio-based BVA method uses a relative entropy metric to determine the relative importance of each stream in a group of signals. We evaluate the proposed method by comparing the value assignments made by our algorithm to those made by human subjects on a small situation awareness task. In the user study, we found that our BVA method performed significantly better than chance and within a small delta of average human performance. Moreover, we found that performance *increases* on those segments which have high levels of human agreement. We compared our audio-based BVA method to video-based systems using color and motion features; our results indicate that auditory information is superior to video information for this task. We thus conclude that our method, which uses inexpensive auditory processing, can be used to enhance the utility of online sensory data in multiparty collaborative tasks.

We discuss related work in Sec. 2 before detailing our proposed method in Sec 3. Our experimental setup is discussed in Sec 4, the results of those experiments are found in Sec 5. Conclusions can be found in Sec 6.

2. RELATED WORK

BVA is closely related to scene change detection. Scene changes in sensory data represent boundaries across which the sensory data changes in some way; standard perceptual theory indicates that boundaries in sensory data are areas of high information content and are in this sense valuable for information extraction. See e.g. [1], [2], [3] for good discussions of these concepts and how they inform the design of automatic sensory processing systems.

Methods for scene change detection in audiovisual data abound. Slaney et. al. proposed the use of temporal derivatives of scale-space audio and video features to find scene changes at multiple timescales in [4] and [5]. Foote proposed a temporal structure feature based on self-similarity for a similar purpose [6], [7]. Gaborski et. al. proposed a method for novelty detection in video [8], [9] which uses a relaxed k-means clustering over contrast, edge, color, and motion features together with a habituation component.

Work on context aqcuisition from real-world sensory data, particularly audio, is a growing field. Important steps toward this goal have been taken by many researchers, including Clarkson ([10], [11]) and Ellis. Of particular interest is the work in [12] and [13], in which real-world audio data are segmented according to a BIC measure (introduced by Chen et. al. in [14]) and clustered according to sym-

This research is supported by the European Commission CHIL project (http://chil.server.de) under contract No. 506909.

metrized relative entropy.

3. BLIND VALUE ASSIGNMENT IN OAS APPLICATIONS

In BVA, we considering statistical measures of information content as proxies for value. The information content of a data stream in isolation can be estimated by calculating its entropy. In media summarization applications, though, simply ranking segments by entropy might yield a summary with many similar segments. This kind of redundancy is usually to be avoided both on theoretical and practical grounds; the proper approach is to maximize the entropy of the summary itself, or equivalently, to minimize the mutual information between summary segments. Minimizing mutual information is in turn equivalent to finding segments which are in some way distant from one another. In this research, we use a normalized, symmetrized version of relative entropy to estimate distance and hence value of one stream relative to some other group of streams. This metric measures the percentage of the coding cost that is wasted when we use a model inferred from one set of data to code a different set of data. Symmetrized relative entropy, D^2, given in Eqn. 1, measures the absolute cost of using the wrong model in this way. As relative entropy can be rewritten as the difference between cross entropy and entropy, dividing the D^2 metric by the sum of cross entropies yields the waste percentage. This final metric, $\hat{D}^2$, is given in Eqn 3.

$$D^2(p||q) = D(p||q) + D(q||p), \quad (1)$$
$$= H_{p,q}(X) + H_{q,p}(X) - H_p(X) - H_q(X). \quad (2)$$

$$\hat{D}^2(p||q) = 2 - \frac{H_p(X)}{H_{p,q}(X)} - \frac{H_q(X)}{H_{q,p}(X)}. \quad (3)$$

In OAS applications, the goal is not to produce a summary; it is to direct attention in real time to the most interesting streams. Hence, our only goal is to compute the value of each stream relative to the rest of the stream set at a given time. There are two ways in which a stream could be valuable relative to the other streams. First, the stream could contain information which is different from all the other streams. We call this feature *uniqueness*. Second, the stream could contain information that is new for that stream; i.e. it could represent a scene change. We call this feature *novelty*. In our experiments, we considered two types of novelty. The first type, which we called *history*, compares a stream segment to the previous three segments of that stream. The second type, which we called *scene change*, estimated with the $\hat{D}^2$ metric the degree to which the scene changed across three different boundaries within the same segment. Intuitively, it is important to consider both uniqueness and novelty when assigning value in the OAS scenario. In this study, we considered linear combinations of uniqueness and novelty, and compared the use of the history and scene change variants of novelty.

4. EXPERIMENTS

We conducted a set of experiments to evaluate the performance of our audio-based BVA method. To carry out these experiments, we recorded several real-world audiovisual streams, arranged them into parallel collections of segments, and estimated the value of each segment. In addition to audio-based judments, we also computed value estimates using video features for comparison. To evaluate performance, we elicited value judgments from humans and used them to score the automatic judgments.

4.1. Data Collection

Our dataset consisted of four short audiovisual recordings made by one of the authors as he performed various errands around the Carngie Mellon University campus. These errands included purchasing lunch from a mobile vendor, visiting an ATM to make a withdrawal, mailing a letter and buying a soda, and filling his car with gasoline. The scenes ranged from 11 to 15 minutes and were made with a Hitachi MPEG-1 video camera mounted, front-facing, to the author's backpack.

4.2. Audio Feature Extraction

Audio was extracted from the MPEG-1 stream and converted into 16-bit, 16kHz raw data. From the raw data, we computed three separate sets of audio features, all at a rate of 100 frames per second. First was a set of 20 melscale frequency coefficients (MEL). The melscale filterbank is a widely-used approximation of the frequency responses of mammalian auditory systems; the bandwidth of each filter is inversely proportional to the center frequency. Second was a set of 20 mel-frequency cepstral coefficients (MFCC), computed by applying a log nonlinearity to melscale filterbank coefficients and applying the discrete cosine transform. MFCCs are the most widely used features in automatic speech recognition and other auditory analysis tasks. Third was a package of spectral summary features (PERC) each designed to highlight some relevant meta-feature of the spectrum. Using the power spectrum as the source, we computed the spectral centroid, the root mean squared energy, the signal-to-noise ratio, and a band energy ratio comparing energy below 200Hz to energy above 200Hz. Spectral centroid is a measure of the perceptual "brightness" of a sound and has been used both for acoustic event classification and scene change detection. Root mean squared energy is a measure of loudness and has been used to detect acoustic events and scene changes. The signal-to-noise ratio measures the noisiness of a sound field and is an important textural feature. Finally, the band energy ratio is designed to detect sound fields dominated by low-frequency noise. In our experiments, we used two merged feature sets: MEL+PERC and MFCC+PERC.

4.3. Video Feature Extraction

We also computed two sets of video features. First was a set of color features (COLOR). This feature set was made up of 128 histogram values in the HSV (Hue, Saturation, Value) color space, sampled at a rate of 30 frames per second. Second was a set of motion history features (MOTION). The motion features consisted of a vector of five values; energy, horizontal motion mean, vertical motion mean, horizontal motion variance, and vertical motion variance. These features are intended to characterize both global camera motion and local object motion. We computed these features based on optical flow outliers every three video frames, resulting in a rate of 10 motion history vectors per second. We used the Lucas-Kanade [15] algorithm to compute the motion features.

4.4. Data Modeling

For all feature sets, we modeled each segment or subsegment with a single multidimensional Gaussian with a diagonal covariance matrix. We elected not to use mixtures of Gaussians for several reasons. Parameter estimation is both faster and more robust for one model than many, given the same amount of input data. Furthermore, our goal was not necessarily to produce accurate models of the data; rather, it was to estimate how much information is present in the

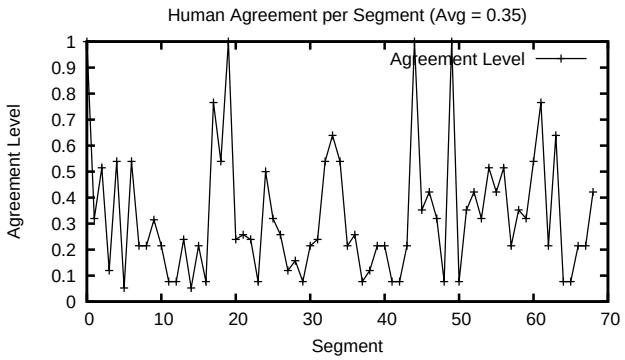

Fig. 1. Agreement coefficients among human subjects per segment

data. In this case, using the covariance matrix of a single Gaussian to measure information content amounts to placing an upper bound on information content, which we viewed to be sufficient for our purposes.

4.5. Eliciting Human Value Judgments

In order to elicit human value judgments, we aligned the four audiovisual streams in time, cutting them to the length of the shortest stream. We then divided the streams into 69 10-second segments. Using the SMIL markup language, we constructed a four-window display to play the video streams in parallel or in sequence at the user's choice. In the parallel case, the audio from all streams was mixed into a single stream. We presented the series of four-stream segments to the subjects, and asked them to rank the streams by value, given that their task was to attain situational awareness over the entire set of activities. For this study, we used 10 human subjects.

We found that in many cases, there was little agreement about which streams were most important. In fact, there were many segments in which not much of anything was happening in any stream; in those cases, subjects simply chose a stream at random or according to some esoteric preference. However, there were a significant number of segments where agreement was greater than one would expect by chance. These segments were mostly those which contained clearly valuable information; e.g. the segment where the ordered his lunch or started pumping gas. We measured human agreement levels by computing the Shannon redundancy of the responses per segment. Shannon redundancy measures the divergence between a discrete distribution and the uniform distribution. A value of 1 indicates unanimity. Human agreements per segment are shown in Fig. 1. The average agreement level was 0.35; in 18 segments the agreement level was 0.5 or higher.

Tab. 1 shows the result of taking each human out of the reference set and computing a match score against the rest of the subjects. Notably, the "best" performing human matched the remaining subjects' responses less than half the time. We take the average human performance level to be a reasonable target for machine performance.

5. RESULTS

We evaluated our audio-based BVA method against human judgments for all feature sets, at several different weightings of uniqueness and novelty, and using both the history and scene change variants of novelty. Match percentages were computed by counting each human judgment match and dividing by the total number of human judgments (690). We also computed restricted match rates, in which we eliminated segments which displayed low human agreement levels. Audio results are discussed in Sec. 5.1, video results in Sec. 5.2, and agreement-restricted results in Sec. 5.3.

Human Performance			
Subject	Match Rate	Subject	Match Rate
h1	42.19%	h6	33.97%
h2	**46.85**%	h7	39.45%
h3	42.35%	h8	42.99%
h4	39.29%	h9	41.22%
h5	42.02%	h10	39.61%
Average			40.99%

Table 1. Human match rates per subject

Nov Wt	Feature			
	MEL		MFCC	
	Nov Type		Nov Type	
	Hist	Scn Chg	Hist	Scn Chg
1	23.53%	**36.32**%	25.44%	33.08%
0.75	25.44%	32.94%	25.58%	30.00%
0.5	23.23%	26.32%	24.11%	20.14%
0.25	20.44%	23.97%	22.05%	20.73%
0	23.08%		20.73%	

Table 2. Audio BVA match rates, MEL and MFCC features

5.1. Audio Results

Audio results are shown in Tab. 2. There are several results of interest in this set of figures. First is that the use of the scene change variant of novelty outperformed the history variant by a significant margin. This is likely due to human inability to remember a long history for each of four streams, while scene changes within a single segment are obvious. Second, novelty appears to be much more relevant than uniqueness. This result squares with standard perceptual theory which holds that boundaries are the most informative regions. Third, the MEL feature set scored 2.5% better absolute than the MFCC feature set. This is likely due to the fact that MFCCs obscure pitch information, which could conceivably be of use in this task. Finally, the optimal result of 36.3% is 11.3% better than chance and 4.7% worse than the average human, meaning that nearly 70% of the difference between these two levels of performance was covered.

5.2. Video Only Condition

Video results are shown in Tab. 3. For the MOTION set, we used only the history variant of novelty, as there were not enough frames per segment (30) to perform scene change detection. Here, we see that in only two conditions did the results exceed chance. Color features were superior to motion features, history was better than scene change, and novelty appeared to be more relevant than uniqueness. We suspect that the video features performed so poorly compared to the audio features for several reasons. First, distance is a factor in measuring changes in color and motion. That is, on object that is moving very close to the camera will cause a radical change in the features, and hence a high information score. These kinds of changes may be spurious with respect to the task. Second, it is diffi-

Nov Wt	Feature		
	COLOR		MOTION
	Nov Type		Nov Type
	Hist	Scn Chg	Hist
1	**26.61**%	16.91%	21.76%
0.75	24.41%	23.38%	18.38%
0.5	26.03%	23.38%	15.58%
0.25	22.94%	23.38%	15.29%
0	23.38%		15.00%

Table 3. Video BVA match rates, COLOR and MOTION features

System	Agreement Level			
	0.25	0.5	0.75	1.0
MEL	48.78%	**61.11**%	48.33%	50.00%
MFCC	39.09%	39.44%	35.00%	50.00%
COLOR	28.18%	33.33%	22.78%	50.00%
MOTION	25.00%	16.67%	21.67%	22.42%

Table 4. BVA match rates at selected human agreement levels

cult to detect certain classes of scene change in video. For instance, if the subject is outside walking towards a door for several seconds, the building will only slowly begin to dominate the visual field; once through the door, there may not be much change at all. In audio, on the other hand, these types of changes are almost always very abrupt and easy to detect.

5.3. Agreement-Restricted Results

Perhaps more important than achieving a high match rate on this task is achieving a high match rate in those cases where human agreement is high. To measure this condition, we considered subsets of segments where human agreement exceeded some threshold. We used four different thresholds for this purpose; the results are summarized in Tab. 4. The audio systems perform better on those subsets of segments in which humans broadly agree. Since there are very few segments at levels 1.0 and 0.75, the performance at level 0.5 is most relevant. Here the MEL system matches human judgment over 60% of the time.

6. CONCLUSION

We have presented a blind audio-based method for assigning value to online aggregate sensor streams with the goal of directing the attention of a planning agent and thus reducing cognitive load in mulitparty, collaborative applications. We evaluated this method by constructing a scenario, eliciting human value judgments on real-world data, and comparing our automatically-derived judgments to human value judgments. We found that our audio-based systems produced value judgments that were broadly similar to human judgments. Specifically, we found that our best system covered nearly 70% of the difference in between chance the average human subject in our experiment. This finding suggests that our system could be usable in real-world applications. We also found that scene change appears to be the most reliable indicator of value. This finding squares with standard perceptual theory, which indicates that sensory boundaries are more informative than non-boundaries. Finally, our audio-based systems performed better than the video-based systems we compared them to, confirming our belief that audio information is better suited for the task of value assignment in real-world data than video information.

7. REFERENCES

[1] J.J. Atick, "Could information theory provide an ecological theory of sensory processing?," *Network: Computation in Neural Systems*, 1992.

[2] A.J. Bell and T.J. Sejnowksi, "The 'indepdendent components' of natural scenes are edge filters," *Vision Research*, 1997.

[3] P. Smaragdis, *Redundancy reduction for computational audition, a unifying approach*, Ph.D. thesis, MIT, 2001.

[4] M. Slaney, D. Ponceleon, and J. Kaufman, "Multimedia edges: Finding hierarchy in all dimensions," in *Proceedings of the Interational Conference on Multimedia and Expo*, 2001.

[5] M. Slaney, D. Ponceleon, and J. Kaufman, "Temporal events in all dimensions and scales," in *Proceedings of the International Conference on Computer Vision*, 2001.

[6] J. Foote, "Visualizing music and audio using self-similarity," in *Proceedings of ACM Multimedia*, 1999.

[7] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proceedings of the International Conference on Multimedia and Expo*, 2000.

[8] A. Vaingankar, V. Chajoi, R. Gaborski, and A. Teredesai, "Cognitively motivated habituation for novelty detection in video," in *NIPS Workshop on Open Challenges in Cognitive Vision*, 2003.

[9] R. Gaborski, "Automatic detection of novel scenes in video," in *University Technology Showcase*, 2005.

[10] B. Clarkson and A. Pentland, "Extracting context from environmental audio," in *Proceedings of International Symposium on Wearable Computers*, 1998.

[11] B. Clarkson, *Life Patterns: Structure from Wearable Sensors*, Ph.D. thesis, MIT, 2002.

[12] D. Ellis and K.S. Lee, "Minimal-impact audio-based personal archives," in *First ACM Workshop on Continuous Archiving and Recording of Personal Experiences*, 2004.

[13] D. Ellis and K.S. Lee, "Features for segmenting and classifying long-duration recordings of personal audio," in *Workshop on Statistical and Perceptual Audio Processing*, 2004.

[14] S. Chen and P. Gopalakrishnan, "Speaker, environment, and channel change detection and clustering via the bayesian information criterion," in *DARPA Speech Recognition Workshop*, 1998.

[15] B. Lucas and T. Kanade, "An iterative registration technique with an application to stereo vision," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 1981.