ABSTRACT

Many critical applications are characterized by groups of humans coordinated by a central planning authority and working towards a common goal. As bandwidth and sensors have become cheaper and more usable in real-world situations, the temptation is to equip each team member with full sensory apparatus and hence create a record which perfectly transmits in real time relevant information about each member’s state and environment. However, as team numbers grow, so does the risk of information overload; it is likely that the value of additional, unfiltered information beyond some optimal point is actually negative.

In this paper, we propose a fast, robust multimodal value assessment technique to help reduce cognitive load. The system automatically selects from multiple sensory streams the single stream judged to be most interesting. The system is fast and robust in that it uses a notion of value based on information theoretic measures of novelty and requires no prior knowledge of situations or environments. We evaluate the system by comparing the system’s value assessments to those of human subjects.

Keywords

Context-aware multimodal processing, ubiquitous and pervasive environments, multimodal-multisensor interfaces, collaborative interaction and systems

1. INTRODUCTION

Many critical applications are characterized by groups of humans coordinated by a central planning authority and working towards a common goal. In that the planning activity depends on accurate context awareness, it is clear that information must be transmitted from team members in the field to the central authority. Given the growing robustness and shrinking expense of sensors and bandwidth, it is tempting to propose that each team member be equipped with a full audiovisual sensor suite capable of constantly transmitting telemetry to the central authority. This approach, while admirable in that it can transmit a more or less complete picture of the activity, is flawed in that the amount of data can quickly become overwhelming; it is likely that once team size grows beyond a few members, cognitive overload would become an impediment rather than a benefit to quick acquisition of context awareness.

To decrease the risk of information overload, we require a means of filtering information such that only the most pertinent data are presented to the planning authority. One such method might be to pre-filter telemetry streams in the field — that is, to only send data from a team member when that member’s situation changes in an observable way; for instance, when a firefighter breaks down a door and finds a burning room. The disadvantages of this method are first that full telemetry is not necessarily available to the central authority if needed (i.e. for after-action review), and second that it fails to account for the situation as a whole. That is, what the central authority considers important information from some team member may change depending on what other team members are experiencing. The alternative is to always transmit telemetry from each member and to select, at the central authority, given all telemetry streams, which are the most interesting. This problem is known as the value assessment problem and is an area of active research; e.g. for personal diary systems, multimedia information retrieval, and the like.

Early approaches to the value assessment problem in the video skimming domain relied on shot detection [9], key frame extraction [10], and mosaic construction [4]. The problem we consider here is more complex in that we are not seeking summaries from a single stream, but valuable segments from multiple streams.

In this paper, we propose a fast and robust multimodal method for selecting interesting audiovisual streams from groups of such streams. The system is fast and robust in that it relies on information-theoretic notions of novelty and interestingness, and requires no prior knowledge or detailed models of expected environmental conditions or event types. We evaluate our method by comparing machine-selected segments to human-selected segments over the same suite of audiovisual streams.

The remainder of this paper is organized as follows. A detailed description of our approach is presented in Section 2. The data we collected to evaluate the approach is presented
in Section 3. Feature extraction is discussed in Section 4 and experimental results are discussed in Section 5. Conclusions are found in Section 6.

## 2. VALUE ASSESSMENT APPROACH

Our online value assessment approach is restricted by the envisioned application to be fast and robust to environmental conditions. Both of these requirements lead us to prefer shallow analysis which does not rely on detailed prior knowledge of typical environments or events. With this preference in mind, we first discuss our modeling framework before presenting our mathematical definitions of value.

### 2.1 Audiovisual Modeling

We model the audiovisual stream as a conjunction of two multidimensional gaussians; one for audio and one for video. The pdf for an n-dimensional gaussian distribution with mean vector $\mu$ and covariance matrix $K$ is given in Equation 1. In practice, we assume decorrelation among dimensions and employ a diagonal covariance matrix.

$$P_{\text{gauss}}(x|\mu, K) = \frac{1}{(2\pi)^{\frac{n}{2}}|K|^\frac{1}{2}}e^{\frac{1}{2}(x-\mu)^T K^{-1}(x-\mu)} \quad (1)$$

### 2.2 Measures of Interest: Novelty and Uniqueness

As noted above, value assessment requires some notion of interestingness. One approach to measuring interestingness might be to make use of entropy (Equation 2), which measures the information content in some random variable $X$.

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \quad (2)$$

Measuring the entropy of an audio stream $X_A$ or video stream $X_V$ is not ideal for this application, as it makes no reference to how much new information is present in the distribution $p(X)$, given that we already have some knowledge about the distribution of some other random variable $Y$ (or, some other distribution $q(\cdot)$ over the same random variable). That is, we cannot use entropy to measure how different two distributions of random variables are. This limitation makes entropy ill-suited for the task of assessing relative value among multiple audiovisual streams.

Information theory does provide a means of measuring the difference between two distributions over a random variable in the form of the Kullback-Liebler Divergence (KL-D) [2], or relative entropy, given in Equation 3.

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (3)$$

Relative entropy can be defined as the inefficiency incurred by assuming that the distribution over some random variable $X$ is $q(\cdot)$ instead of the true distribution, $p(\cdot)$. Though relative entropy is not symmetric, it is typically informally used as a distance metric between distributions.

Relative entropy can be rewritten as the difference between two terms, as shown in Equations 4 and 4. These terms are the cross entropy $CH(p,q)$, shown in Equation 5, and the entropy $H(p)$. As this expansion implies, a major weakness of using relative entropy for this application is that segments are penalized for having high entropy. As we equate high-entropy segments with segments of interest, this is an undesirable result. We can, however, use the cross entropy alone to measure interest. Given the intuitive definitions of relative entropy (i.e. the inefficiency in assuming an incorrect distribution) and entropy (i.e. the inherent cost to encode a random variable), we can define cross entropy as the total cost of using an incorrect distribution to represent a random variable. This means that cross entropy is potentially a better measure than relative entropy for this application.

$$D(p||q) = \sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} p(x) \log q(x) \quad (4)$$

$$CH(p,q) = -\sum_{x \in X} p(x) \log q(x) \quad (5)$$

In this application, we can use $p(X)$ to represent the distribution over audio or video in a given stream segment whose value we are assessing, and $q(X)$ to represent the distribution over audio or video in the remaining streams that serve as the reference model. In this application, $p(X_t)$ would represent a single stream segment at time index $t$. To measure uniqueness, $q(X_t)$ would represent the other stream segments at the same time index. To measure novelty, $q(X_{t-1}, \ldots)$ would represent the same stream segment over some previous time window.

### 2.3 Approximating Cross Entropy

A closed form for the cross entropy between two gaussian distributions $p(X)$ and $q(X)$ is given in Equation 6.

$$CH(p,q) = \frac{1}{2} \log(2\pi)^n |K_q| + \frac{1}{2 |K_p|} [||K_p| + (\mu_p - \mu_q)^2] \quad (6)$$

In this application, we often lack the requisite number of samples to properly estimate the parameters of even a single multidimensional gaussian for the test segment which is modeled by $p(X)$ (whereas we can have many more samples to estimate the reference model $q(X)$). Alternatively, if we view cross entropy as the average surprise of the model $q(X)$ when exposed to samples from $X$, then we have:

$$CH(p,q) \approx -\sum_{x \in X} \log q(x) \quad (7)$$

This approximation is equivalent to the opposite of the average log probability of $X$ given the model $q$: $\mathcal{L}(X|q) = \sum_{x \in X} \log q(x)$ (8)

Hence, we substitute $\mathcal{L}(X|q)$ for $CH(p,q)$.

## 3. DATA COLLECTION

We collected four data streams with which to evaluate our approach. We used an Hitachi model MP-EG10W portable mp3g camera. The camera captured MPEG-1 video at a rate of 30 frames per second and 16-bit audio at 32kHz. The camera was mounted on a backpack strap and carried by one of the authors as he carried out four simple tasks:...
getting cash from an ATM (Figure 1), mailing a package (Figure 2) and purchasing a cola from a machine (Figure 3), buying lunch from a mobile vendor (Figure 4), and putting gasoline in his car (Figure 5). In all four scenes, the subject started in his office, walked out, performed the task, and returned to the office. The four scenes varied in length from 688 seconds to 914 seconds.

4. FEATURE EXTRACTION

We extracted both audio and video features from the captured data streams. We describe the feature extraction processes for audio and video in turn.

4.1 Audio Features

Audio is downsampled from 32kHz to 16Khz with 16-bit depth. We compute two sets of features from the raw audio; a spectral feature set (SPEC), intended to capture and a perceptual feature set (PERC). For the spectral feature set, we first compute a Short-Time Fourier Transform (STFT) from the raw audio with a 200msec sample window and 100msec overlap, resulting in 10 frames per second. We pass the 2049-point STFT power spectrum through a 20-dimensional mel-scale filterbank and then apply the log. These 20 log-melscale features (LMEL) make up the spectral feature set. The perceptual feature set is made up of eleven features, each intended to capture some perceptually relevant feature of the audio that is not well-represented by the spectral features. Included in this set of perceptual features are five Band Energy Ratios (BER, Section 4.1.1), Bandwidth (BW, Section 4.1.2), Power (POW, Section 4.1.3), Log Zero-Crossing Rate (LZCR, Section 4.1.4), Spectral Division (DIFF, Section 4.1.5), Brightness (BRT, Section 4.1.6), and Significant Subband Count (SSC, Section 4.1.7). These features, like LMEL, are computed at the rate of 10 frames per second. The LMEL and perceptual feature sets are merged to comprise the audio feature set used for value assessment.

4.1.1 Band Energy Ratio

Band Energy Ratio is the ratio between spectral energy below a given threshold frequency \( f \) and spectral energy above that frequency in an analysis frame. In this study, we compute BERs at 50Hz, 500Hz, 1kHz, 2kHz, and 3kHz. The BER is computed as in Equation 9. Here, and in the following equations, \( S_i^t \) refers to the spectral energy in the STFT at the \( i^{th} \) frequency of the analysis window \( t \), and \( f_i \) refers to the \( i^{th} \) frequency.

\[
BER_i^t = \frac{\sum_{j=0}^{i} (S_j^t)^2}{\sum_{k=i+1}^{N} (S_k^t)^2}
\]  

(9)

4.1.2 Bandwidth

Bandwidth is the difference between the highest frequency whose energy exceeds some threshold \( \alpha \) and the lowest frequency whose energy exceeds \( \alpha \). BW is computed as in Equa-
 convention.

\[ BW_t^a = \arg_{t_i} \max[S_t^i > \alpha] - \arg_{t_i} \min[S_t^i > \alpha] \]  

(10)

4.1.3 Power

Signal Power is the average squared magnitude of the raw audio samples over some analysis window and is computed as in Equation 11. Here, \( A_t^j \) refers to the \( j \)th raw audio value in analysis window \( t \).

\[ \text{POW}_t = \frac{\sum_{j=0}^{M} (A_t^j)^2}{M} \]  

(11)

4.1.4 Log Zero Crossing Rate

Zero Crossing Rate is the count of instances of sign switching from one raw audio sample to the next over some analysis window \( t \). Here, we use the Log Zero Crossing Rate, which is computed as in Equation 12.

\[ \text{LZCR}_t = \log_2 \left( 1.0 + \sum_{i=1}^{M} \left\{ \begin{array}{ll} 1 & \text{if } A_t^i A_{t-1}^i < 0 \\ 0 & \text{otherwise} \end{array} \right. \right) \]  

(12)

4.1.5 Spectral Diffusion

Spectral Diffusion is an entropy-like feature which measures the spread of spectral energy across frequencies. After normalizing the STFT to give it the appearance of a pdf as in Equation 14, we compute DIFF as in Equation 13.

\[ \text{DIFF}_t = -\sum_{i=0}^{N} \hat{S}_t^i \log \hat{S}_t^i \]  

(13)

\[ \hat{S}_t^i = \frac{S_t^i}{\sum_{i=0}^{N} S_t^i} \]  

(14)

4.1.6 Brightness

Brightness, or spectral centroid, is the mean spectral energy weighted by frequency. BRT is computed as in Equation 15.

\[ \text{BRT}_t = \sum_{i=0}^{N} f_i S_t^i \]  

(15)

4.1.7 Significant Subband Count

The Significant Subband Count is the number of frequencies in an analysis window whose energies exceed some threshold \( \alpha \). SSC is computed as in Equation 16.

\[ \text{SSC}_t^\alpha = \sum_{i=0}^{N} \left\{ \begin{array}{ll} 1 & \text{if } S_t^i > \alpha \\ 0 & \text{otherwise} \end{array} \right. \]  

(16)

4.2 Video Features

The most common feature sets used in video and image indexing and retrieval are motion features, shape features, texture features and color features. Motion features are crucial for characterizing motion patterns of objects in video; for example the Motion History Image and optical flow. Shape features are employed mostly in object retrieval tasks to provide identification information of specific objects; for example Hough transform parameters and elastic model parameters. Simple shape features are also developed to distinguish individuals from groups of objects [1]. Extraction of motion and shape features is computationally expensive, especially for videos, like the ones described in this work, which contain a lot of motion. Texture features provide information when color is not distinguishable or available. Color features have proven to be the most robust type of feature for video skimming on top of text descriptions. Previous results have shown that representations based on regional or global color histograms [5], [7], [8] are fast and efficient ways to characterize video content.

In this paper, we use color features only. We intend to make use of motion features in future work. In the following, we discuss relevant motion features before describing the color features that we use in this study.

4.2.1 Motion vector

The most common motion features are motion vectors of macro-blocks in the mpeg and H.26x codecs.

Formally, let us denote the vector \( v = (v_x, v_y) \) as the motion vector pointing the block \( B \) to its matching block in the reference image. The optimal motion vector \( v^* \) can be computed by minimizing the sum-of-squared difference over all the pixels \( x \in B \) in the block \( B \):

\[ v^* = \arg \max_{v=(v_x,v_y)} \sum_{x \in B} [I_t(x) - I_{t-1}(x-v)]^2, \]  

(17)

where \( I_t(x) \) indicates the intensity value of the pixel \( x \) in the current image and \( I_{t-1}(x-v) \) represents the intensity value of the pixel \( x \) shifted by the vector \( v \) in the reference image.

4.2.2 Image differencing

A simple way to produce dense motion feature is to compute the difference between two consecutive images. For any given frame of image \( I_t \) at time \( t \) and its previous image \( I_{t-1} \) at time \( t-1 \), the difference motion feature at each pixel \( x \) is computed as:

\[ D_t(x) = I_t(x) - I_{t-1}(x). \]  

(18)

The motion features of all pixels can be represented as a differencing image \( D_t \). This feature has a serious drawback because the value of \( D_t(x) \) indicates neither the direction of the movement of the pixel \( x \) nor the velocity value.

4.2.3 Optical flow

An alternative approach for obtaining dense motion features is optical flow. Optical flow techniques compute the velocity of each pixel, feature points or region from spatiotemporal derivations of image intensities. These methods assume the intensity of each pixel is conserved temporally and the changes between consecutive images are caused by movements. Formally, a frame of image at time \( t \) can be represented as a spatiotemporal function \( I(x,t) \):

\[ I(x,t) = I(x-vt,0), \]  

(19)

where \( v \) is the velocity. Imposing the temporal conservation assumption on the Taylor expansion of Equation 19, the gradient constraint equation is derived as:

\[ \nabla I(x,t) \cdot v + I_t(x,t) = 0, \]  

(20)
where $I_t(x, t)$ denotes the temporal partial derivative of $I(x, t)$ and $\nabla I(x, t)$ denotes the spatial partial derivatives of $I(x, t)$ in horizontal and vertical directions. To solve the Equation 20, further constraints need to be introduced because the velocity $v$ has two unknown components in both horizontal and vertical directions.

An example is the Lucas-Kanade (LK) [6] algorithm, which constrains the velocity $v$ to be a constant vector in each small spatial neighborhood $w$ and optimizes velocity $v$ by minimizing a weighted sum-of-squares of Equation 20:

$$v_w = \arg \max_v \sum_{x \in w} W(x) \left( \nabla I(x, t) \cdot v + I_t(x, t) \right)^2, \quad (21)$$

where $W(x)$ are constant weights that decrease from the center of the neighborhood $w$.

To obtain good optical flow, a video sequence is usually smoothed by applying a low-pass spatiotemporal Gaussian filter to help attenuate temporal aliasing. Optical flow can provide dense motion feature even for every pixel. However, the computation cost of the extraction of optical flow is high. Therefore, only the first-order based methods, such as the LK method, have potential applications in analyzing large amounts of video data.

### 4.2.4 Edge motion history image

Motion history image (MHI) was proposed by [3] to analyze human activities. An MHI is a compact temporal template which represents recent object movements. In [3] an MHI is computed from silhouettes of objects segmented using background subtraction and stereo depth subtraction. The background is not easy to extracted in news and sports videos with complex background scenes. Further, stereo depth information is usually not available in video.

Let $E_t(x)$ be a binary value to indicate if pixel $x$ is located on an edge at time $t$. A EMHI $H^E_t(x)$ is computed from the EMHI of the previous frame $H^E_{t-1}(x)$ as:

$$H^E_t(x) = \left\{ \begin{array}{ll} \tau, & \text{if } E_t(x) = 1 \\ \max(0, H^E_{t-1}(x) - 1), & \text{otherwise}. \end{array} \right. \quad (22)$$

Computation of EMHI is less expensive than computation of optical flow, making it a better candidate for realtime video analysis.

### 4.2.5 Color Features

In this paper, our video features are global color histograms over each frame. Specifically, we use the HSV (Hue, Saturation, Value) color space rather than the RGB color space. The HSV color space provides a potentially better feature space than the RGB space because hue is much more representative than saturation and value. To compute the color features, we first convert the raw RGB video images into the HSV color space and then quantize the HSV into bins. The color histogram is then defined as the joint distribution of the intensities of the three quantized color channels. In this study, we compute 128 histogram bin values at the rate of 30 frames per second.

5. EXPERIMENTS

We evaluated the performance of our value assessment method by comparing machine-assigned values to human-assigned values over the audiovisual data we collected. Humani-
assigned values were collected by presenting parallel segments from each audiovisual stream to human subjects and asking them to rank the streams from most interesting to least interesting. Definition of “interesting” was intentionally left vague in our instructions. The subjects were, however, apprised that the task domain was situation awareness for activity planning.

We limited our evaluation to the first 690 seconds of each stream, which is the length of the shortest stream. We first divided the streams into 3 230-second substreams, and assigned seven humans to each substream. We then divided each substream into 23 10-second chunks for evaluation. We presented the humans with a four-stream display for each 10-second chunk, as shown in Figure 9. We used the rankings collected in this manner to evaluate our value assessment method.

5.1 Audio-Audio, Video-Video, Audio-Video Agreement

5.2 Human-Machine Agreement

5.2.1 Audio-Only Condition

5.2.2 Audio + Video Condition

<table>
<thead>
<tr>
<th>KL Divergence Video-Video Agreement</th>
<th>Uniqueness Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00      1.000     0.895     0.835     0.746     0.432</td>
<td></td>
</tr>
<tr>
<td>0.25      1.000     0.940     0.850     0.492     0.298</td>
<td></td>
</tr>
<tr>
<td>0.50      1.000     0.910     0.552     0.283     0.298</td>
<td></td>
</tr>
<tr>
<td>0.75      1.000     0.626     0.328     0.283     0.283</td>
<td></td>
</tr>
<tr>
<td>1.00      1.000     0.358     0.388     0.343     0.268</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Top-1 agreement among KL-D video systems with different uniqueness weights

<table>
<thead>
<tr>
<th>KL Divergence Audio-Audio Agreement</th>
<th>Uniqueness Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Uniqueness Weight</td>
<td></td>
</tr>
<tr>
<td>0.00      0.388     0.358     0.388     0.343     0.298</td>
<td></td>
</tr>
<tr>
<td>0.25      0.403     0.403     0.432     0.388     0.298</td>
<td></td>
</tr>
<tr>
<td>0.50      0.438     0.473     0.402     0.368     0.283</td>
<td></td>
</tr>
<tr>
<td>0.75      0.428     0.443     0.473     0.328     0.283</td>
<td></td>
</tr>
<tr>
<td>1.00      0.343     0.358     0.388     0.343     0.268</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Top-1 agreement among KL-D audio and video systems with different uniqueness weights

<table>
<thead>
<tr>
<th>Cross Entropy Audio-Audio Agreement</th>
<th>Uniqueness Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00      1.000     0.791     0.656     0.597     0.403</td>
<td></td>
</tr>
<tr>
<td>0.25      1.000     0.865     0.776     0.552     0.403</td>
<td></td>
</tr>
<tr>
<td>0.50      1.000     0.880     0.641     0.403     0.403</td>
<td></td>
</tr>
<tr>
<td>0.75      1.000     0.761     0.328     0.283     0.283</td>
<td></td>
</tr>
<tr>
<td>1.00      1.000     0.761     0.328     0.283     0.283</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Top-1 agreement among CH audio systems with different uniqueness weights

<table>
<thead>
<tr>
<th>Cross Entropy Video-Video Agreement</th>
<th>Uniqueness Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00      1.000     0.825     0.835     0.746     0.432</td>
<td></td>
</tr>
<tr>
<td>0.25      1.000     0.940     0.850     0.492     0.298</td>
<td></td>
</tr>
<tr>
<td>0.50      1.000     0.910     0.552     0.283     0.298</td>
<td></td>
</tr>
<tr>
<td>0.75      1.000     0.626     0.328     0.283     0.283</td>
<td></td>
</tr>
<tr>
<td>1.00      1.000     0.358     0.388     0.343     0.268</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Top-1 agreement among CH video systems with different uniqueness weights
Cross Entropy Audio-Video Agreement

<table>
<thead>
<tr>
<th>Video Uniqueness Weight</th>
<th>0.00</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Uniqueness Weight</td>
<td>0.00</td>
<td>0.313</td>
<td>0.283</td>
<td>0.313</td>
<td>0.268</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.268</td>
<td>0.283</td>
<td>0.313</td>
<td>0.223</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.268</td>
<td>0.283</td>
<td>0.313</td>
<td>0.298</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.253</td>
<td>0.283</td>
<td>0.313</td>
<td>0.238</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.194</td>
<td>0.223</td>
<td>0.238</td>
<td>0.208</td>
</tr>
</tbody>
</table>

Table 6: Top-1 agreement among CH audio and video systems with different uniqueness weights

Table 7: Results, Audio-Only

<table>
<thead>
<tr>
<th>Novelty Wt.</th>
<th>Uniqueness Wt.</th>
<th>Min</th>
<th>Max</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Results, Audio + Video

Table 9: Human-Human Top-1 Agreement

<table>
<thead>
<tr>
<th>Subset</th>
<th>Mean</th>
<th>StdDev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set0</td>
<td>0.496</td>
<td>0.034</td>
<td>0.454</td>
<td>0.565</td>
</tr>
<tr>
<td>Set1</td>
<td>0.319</td>
<td>0.071</td>
<td>0.174</td>
<td>0.435</td>
</tr>
<tr>
<td>Set2</td>
<td>0.352</td>
<td>0.211</td>
<td>0.154</td>
<td>0.695</td>
</tr>
<tr>
<td>Total</td>
<td>0.319</td>
<td>0.188</td>
<td>0.154</td>
<td>0.695</td>
</tr>
</tbody>
</table>

5.3 Inter-Subject Agreement

6. CONCLUSIONS

7. ACKNOWLEDGEMENTS

xxx and xxx are supported by the xxxxxx project under contract No. xxxxxx. xxxxxx is supported by the xxxxxxx project under contract No. xxx. All statements, opinions, and conclusions are those of the authors and do not necessarily reflect the views of the sponsors.

8. REFERENCES


