Multimedia Content Creation using Societal-Scale Ubiquitous Camera Networks and Human-Centric Wearable Sensing

Mathew Laibowitz
Responsive Environments Group
MIT Media Lab
mat@media.mit.edu

Nan-wei Gong
Responsive Environments Group
MIT Media Lab
nanwei@media.mit.edu

Joseph A. Paradiso
Responsive Environments Group
MIT Media Lab
joep@media.mit.edu

ABSTRACT

We present a novel approach to the creation of user-generated, documentary video using a distributed network of sensor-enabled video cameras and wearable on-body sensor devices. The wearable sensors are used to identify the subjects in view of the camera system and label the captured video with real-time human-centric social and physical behavioral information. With these labels, massive amounts of continually recorded video can be browsed, searched, and automatically stitched into cohesive multimedia content. This system enables naturally occurring human behavior to drive and control a multimedia content creation system in order to create video output that is understandable, informative, and/or enjoyable to its human audience. The collected sensor data is further utilized to enhance the created multimedia content such as by using the data to edit and/or generate audio score, determine appropriate pacing of edits, and control the length and type of audio and video transitions directly from the content of the captured media. We present the design of the platform, the design of the multimedia content creation application, and the evaluated results from several live runs of the complete system.

Categories and Subject Descriptors

H.1.2 [**Information Systems Applications**]: Models and PrinciplesUser/Machine Systems[Human Factors, Human Information Processing, Software Psychology]; H.3.1 [**Information Systems Applications**]: Information Storage and Retrieval-Content Analysis and Indexing; H.5.1 [**Information Systems Applications**]: Information Interfaces and PresentationMultimedia Information Systems[Video]

General Terms

Human Factors, Design, Management, Measurement, Theory

MM'10, October 25–29, 2010, Firenze, Italy.

Keywords

Distributed Sensor Networks, Distributed Video Networks, Wearable Sensing, Sensor-Controlled Video Creation, Documentary, User-generated Content, Narrative

1. INTRODUCTION

In this paper we present the SPINNER research platform and its use for the creation of personalized documentary multimedia content. Unlike traditional forms of documenting one's life, the SPINNER system is designed to require no specialized training or skills, yet allows for the creation of comprehensible content. The SPINNER system achieves this goal through the integration of on-body sensor devices with environmentally situated cameras. This allows a participant's behavior, as detected by the sensors, to control the video camera system without the participant changing their activity or perspective in order to capture the event.

More specifically, media and sensor signals are captured by a distributed sensor network equipped with high-quality video capture and a suite of sensate, communicative, and computational facilities. These signals are labeled and constrained from parameters that are collected and broadcast by wearable systems from which location, personal affect, and social parameters can be derived.

This paper summarizes the developed technology of the SPINNER system and presents its deployment and use in a real-life scenario to automatically create several videos. These videos are presented and evaluated to illustrate the effectiveness of the overall system and the developed media creation techniques. More detailed technical information, additional collected data, and descriptions of other applications of the SPINNER system are available in [1].

A system of video cameras deployed in the environment has significant impact on the space, particularly with regards to privacy. Detailed information about the privacy studies and mechanisms deployed to minimize the negative impact of the SPINNER system and future systems of this sort has been previous published in [2].

The SPINNER system addresses the human resource cost associated with multimedia content creation. While the cost of video creation technological is rapidly decreasing, the human-hours required to create cohesive video content has remained relatively constant. By automating this process, the SPINNER system enables considerable resource savings, and this becomes one of the points of evaluation of the system.

2. MOTIVATION

Technological advances such as cheap cameras have shifted the media power structure, allowing mass social engagement in participatory video creation. Media capture technology is exploding, and image/video/audio acquisition capability is pervading many of our common digital devices (e.g., laptops, mobile phones, appliances, toys, etc.), which are increasingly carried on our person, providing opportunities for even more personal media generation. The barriers to entry for contribution of video are non-existent allowing everyone to be a media participant with a voice and a means to document their individual life.

The in-pocket and personally owned form factor of mass-available camera technology promotes image capture based on reaction, not on learned deliberate action. This limits the reach and significance of the captured image, for the most part, to the active participants involved in the original event. In other words, much of the content created in this way is intended as prosthetic memory with nostalgic/informative value only to those immediately involved. To the rest of the world, this appears as noise, often with no way of filtering it out.

The SPINNER system is intended as a platform for experimentation with techniques to take advantage of the new capabilities of partipartory media creation and address the challenges implicit in organizing burgeoning streams of heterogeneous digitized personal media.

3. BACKGROUND AND RELATED WORK

The overall research program presented in this paper lives in a novel and complex application space supported by a combination of a situated distributed video network and wearable sensing that has little to no direct precedence. However, there has been significant work in the related fields from which sub-components of this project have grown. The following subsections illustrate related work in each of these sub-components from which the SPINNER system has based the development of its individual elements to create the overall system in its novel application space.

3.1 Video Sensor Networks

Advances in microelectronics have lead to smaller, cheaper sensor nodes [3], which now sometimes feature video capture. The SenseCam, developed at Microsoft research [4] and now being produced by Vicon [5], brings video sensing and sensor-driven image gathering to a small, low-power node. Devices such as this have began to support a number of distributed camera systems, such as those developed by Wayne Wolf at Princeton as a test platform for distributed vision algorithms [6]. The Panoptes system [9] developed at OHSU demonstrated a reprogrammable video platform and showed how redundancy in smart camera systems can keep information detail even in the event of a network outage.

The work of Andreas Savvides' lab at Yale, called Enalab, often uses multiple camera systems to investigate the real-time capture of diverse phenomena [7], and often uses additional sensing capabilities for subject identification and labeling of video [8].

While not exactly a distributed video network, Microsoft Research's MyLifeBits [10] works to address challenges implicit in organizing burgeoning streams of heterogeneous digitized personal media.

The European Research Council has started an effort known as 4DVideo [11] which seeks to develop algorithms for distributed camera systems to capture and analyze events.

Systems of this type could be integrated into a system like the SPINNER video network and our analysis framework, however, they differ greatly due to their lack of on-body sensing and the subject access that it provides.

3.2 Human-Centric Sensing

Advances in ubiquitous and wearable sensing have allowed us to observe human subjects with minimal interference, supporting research into organizational behavior, social networking, and information diffusion. The Human Dynamics Group at the MIT Media Lab has been a leader in this area through several projects that look at human social behavior, such as the Sociometer [12] and the Reality Mining project [13]. Recent collaborations with the Responsive Environments group have utilized the Uber-Badge platform to look at group social signaling and interest [14]. This has led to a subsequent platform called the Wireless Communicator [15].

Research into the mapping of sensor data to human behavior is exemplified by projects such as Singh's LifeNet [16], Weld's work in personalized user interfaces [17], and Wolf's work on understanding the purpose of travel from GPS data [18]. More specifically, the field of activity recognition from wearable sensors, such as in the work of Stephen Intille at MIT [19] that can classify activity from accelerometer data to provide context in healthcare applications and Paul Luckowicz's work with ETH and the University of Passau on activity recognition using a suite of wearable devices [20], has provided insight for the design of new wearable systems.

In Bove and Mallett's work, a close relative of the SPINNER system because of their attempt to identify knowledge in the data collected from reality, they explore "decentralized approaches for gathering knowledge from sensing devices" [23], including the apropos Two Eyes camera project [24].

4. SYSTEM DESIGN

Our system positions its capabilities where they can be the most effective. The wearable devices are tightly integrated with their subject, allowing access to data pertaining to ID, affective state, social behavior, and human gestural motion. The networked devices in the surrounding environment provide features that require too much power to be included in the wearable suite, are location/environment specific, or otherwise benefit from an objective perspective. Although audio capture can come directly from our electronic badges and some experiments also exploit first-person-perspective video from high-quality cell phone cameras (like that on the Nokia N95) worn around the neck as in [25], the main infrastructure for capturing media are the wall-mounted, sensor-rich Ubiquitous Sensor Portals (USPs).

4.1 Ubiquitous Sensor Portals

The heart of our multi-tier platform is the SPINNER distributed media and sensor system. It integrates into the environment and provides situated, always-on, high-powered, high-bandwidth capabilities to any application. The current deployment has 50 individual nodes, termed Ubiquitous Media/Sensor Portals, installed throughout The Media Laboratory complex at the Massachusetts Institute of Technology.

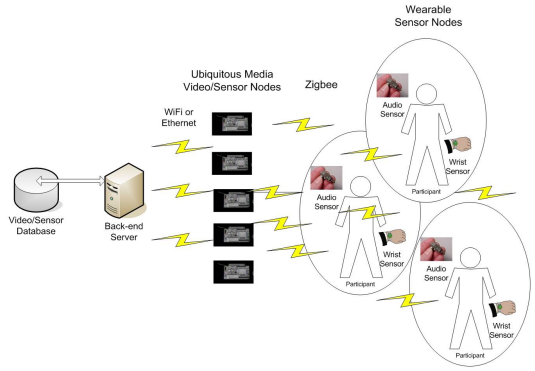

Figure 1: Multi-tiered Distributed Sensor Network

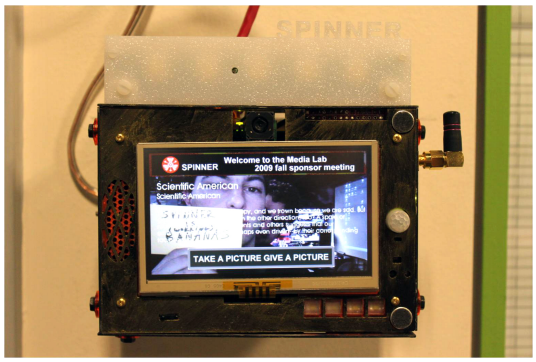

Figure 2: Ubiquitous Sensor/Media Portal

The first component that makes up a USP is the Red Board which is equipped with basic environmental sensors, an audio system with dual powered microphones, and the custom-designed ZMat wireless transceiver module.

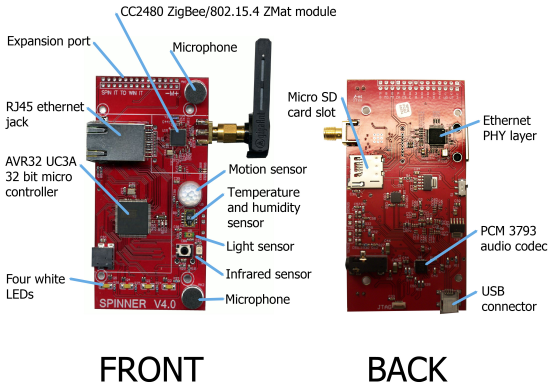

Figure 3: Sensor Node from Distributed Sensor Network, "The Red Board"

The ZMat wireless transceiver module forms a mesh network with all of the always-on and listening Red Boards. This network allows all devices in the system to communicate peer-to-peer as well as to centralized services. Since the Red Boards are also equipped with Ethernet, the wireless transceiver provides internet and high-bandwidth network services to mobile, battery-powered devices that only have a low-powered wireless transceiver.

The next module that the system uses provides media features, such as video capture, and the processing capabilities to handle high-bandwidth data. This module runs an embedded Linux operating system to take advantage of existing development tools and open-source software. This module, nicknamed The Green Board, developed in collaboration with Empower Technologies, is designed to minimize its footprint to aid in deployment, operates with precision timing and real-time features, and provides a suite of hardware optimizations for multimedia capture, image and video processing, and display.

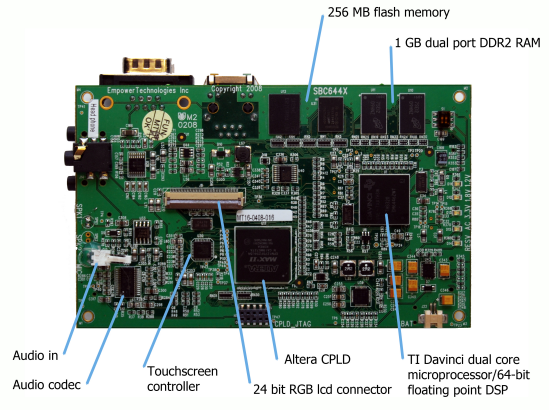

Figure 4: Embedded Audio Video Capture and Processing Device, "The Green Board" (front)

The camera module has a 3.1 megapixel CMOS sensor chip and a motorized auto-focus feature. Each Portal is equipped with a touchscreen for in-situ interactions and distributed media broadcast applications.

An additional module provides high-brightness subject illumination and the control of powerful DC motors used to pan the camera back and forth and tilt then entire unit up and down. This module is known as the Hat, since it normally sits on top of the video camera device. The illumination is provided by 5-Watt LEDs and driving circuitry. It is also equipped with a photographic-quality light sensor.

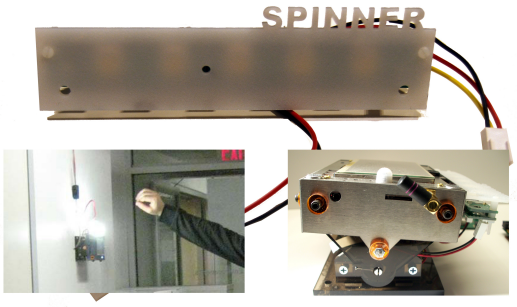

Figure 5: "The Hat" Unit, In Action, Side View of Portal Showing Motorized Mount

4.2 Wearable Sensor Devices

To capture the data most pertinent to the detection of human behavioral events and context, a system of wearable sensors was developed. These on-body sensing devices provide direct access to the subjects and can identify a participant to the system with location information.

The combination of wearable sensing with a distributed sensor network situated in the environment is one of the major achievements of this research. Besides capturing the behavioral data only accessible with wearable sensors, the on-body devices augment the overall network by recording directly from the participant subjective media such as close-mic'ed audio that can be synchronized with media captured throughout the system.

Figure 6: Wearable Sensor Devices

To support a wide range of applications in wearable sensing, two such devices were developed for this research program. These devices, detailed below, are called the SPINNER microBadge, or μBadge, and the SPINNER wrist sensor unit.

4.2.1 *The μBadge*

The first wireless wearable sensor device that we developed is in the form of a small ID badge. This badge is attached to the torso of the wearer either by a lanyard, by a pin, or by being placed in a shirt pocket.

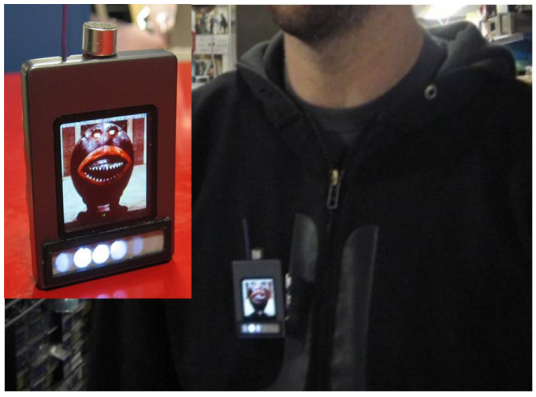

Figure 7: μBadge Wearable Device Pinned to a User

The μBadge builds upon existing badge platforms to investigate conversational dynamics and social signaling [14, 15]. The μBadge has a suite of motion and orientation sensors including a 2-Axis Gyroscope angular rate sensor, 3-axis accelerometer, and a 3-axis tilt compensated compass for world orientation. It is integrated with the SPINNER media network through a ZMat RF transceiver. It is equipped with a dedicated audio CODEC and DSP for capturing and analyzing the wearer's speech in real-time. It is further equipped with removable memory that can store up to two weeks of audio and data, IR transceiver for line-of-sight communication, user-input switches, speaker, and an OLED display.

The ZMat RF transceiver in the wearable devices communicates with the SPINNER network in order to identify and track the participant and to synchronize the data collected from the wearable sensors with that from the environment. In addition to this integration with the larger system, the RF transceiver can be used to off-load recorded data, transmit real-time information, stream audio, and update the device's firmware.

4.2.2 *Wrist Sensor Unit*

The wrist sensor unit is intended to add to the system gestural and biometric sensing channels that can not be seen from the torso-mounted badge. By positioning additional sensors on the wrist, the system can now react to manipulator signaling, specific gestures, context indicated by arm activity, and galvanic skin response.

Figure 8: The Wrist Sensor Unit

The wrist sensor unit is based on the μBadge and contains the same ZMat transceiver module to integrate it with the SPINNER network. The wrist sensor unit has all the same sensors and features as the μBadge, with two major exceptions. The 3-axis accelerometer on the wrist unit is a high-end factory calibrated device providing more accurate and faster sampling for gestural recognition, and the wrist sensor unit has an additional galvanic skin response sensor that responds to user affect.

5. SENSOR DATA FOR VIDEO CREATION

Using candid human physical and social actions as an input to a creative process bases the output of the process on reality, ensuring its validity and potential for relevance. Although reality provides this potential, due to the sheer volume of raw activity (much of which provides no great insight) any actual relevant actions can easily be missed or obscured by overwhelming amounts of non-essential happenings. Having both an always-on distributed video network

and devices directly worn by the subjects give us not only the greatest chance of capturing important events, but also facilitates the tools to understand, label, catalog, and utilize the captured information. The SPINNER Video Application uses the collected sensor data to form and label video segments, which it can then sequence into a cohesive video.

5.1 Video Segmentation

The first task of the SPINNER Video application is to segment the video streams into manageable clips. The wearable sensor devices identify the participants to the system and the location system combined with the Portals' motion detection, both PIR and video, can initially categorize/segment the clips by the subject of clip. The clip, in principle, can immediately be played back on the portal or on the display on the capture participant's wrist sensor unit. Such capability will allow the owner of the clip to preview the clip and then decided if they want to keep it, delete it, or and/or share it.

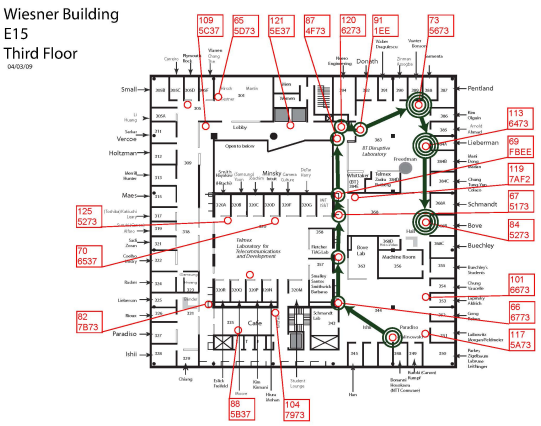

Figure 9: Trajectory of Participant relative to cameras during 1-hour test run

5.2 Social Situation Detection

The clips are further divided according to whether or not a social interaction is shown. This is important information as the wearable sensor data is analyzed differently during a social interaction than during a non-social activity.

The wearable devices use an infrared transmitter and receiver with an adjustable viewing angle to detect when two badges are face-to-face and close enough to indicate a social interaction. This system was first developed for the UbER-badge project in 2004 [28] and has undergone extensive development and testing. The audio system on the μBadge is used to detect who is speaking and who listening. This can be used to analyze the conversation flow, but is also used in conjunction with the wearable sensing, as the sensed parameters also mean different things during speaking and listening.

5.3 Basic Labeling of Video Clips

Once the system has captured and segmented a video clip, it is labeled with some basic information quickly received from the system. These basic parameters are location (which camera it was from), characters (the main character that the clip was based on, as well as any other characters that happened to be in the clip), time (synchronized network time), light level, temperature, humidity, motion (the motion content from the PIR and video), social/not social (whether or not the clip is considered a social encounter or not), and audio (the audio captured from all active badges is synchronized with the video).

The clip detection/segmentation system was evaluated by comparing its results to a human performing the same actions. After a 45-minute run of the system with seven participants, all collected video data was handed off to a non-participant to segment and annotate. The SPINNER Video Application identified and produced a list of 402 video segments, with labels and synchronized audio, while the human identified and produced a list of 396 video segments, a difference of just over 1%.

It took over thirty hours for the human to complete this task, as compared to less than 2 minutes for the SPINNER software. The human could not label the clips with the additional sensor information, and the clip segmentation timings are less precise. This illustrates the utility of the SPINNER Video Application simply as a video annotation tool with enough savings in human resource time to promote the development of large distributed "socially-aware" video camera systems.

The system then labels the video clips with more detailed information from the processing of the wearable sensor data.

5.4 Wearable Sensor Data Processing

Pixel-processing-based vision systems and verbal natural language analysis require a level of domain control over the capture of the audio and video not easily achieved with in-situ camera networks focused on emergent human behaviors. Vision and speech analysis, regardless of whether performed by humans or by machines watching such titanic amounts of video content, are resource heavy and often cannot run in real-time, making them of little use in a large-scale pervasive video capture system.

Perhaps the biggest drawback of video analytics is their limited access to the subject. The cameras in the system provide a remote, third person subjective view, as opposed to an egocentric, body-mounted objective view which, considering the humanity of the subjects, is where the action is.

For the SPINNER system, wearable sensor devices are used to capture on-body low-level sensor data that can be mapped to higher level descriptions of human behavior pertinent to understanding the content of a video clip in the context of the surrounding events and environment. The mapping of raw sensor data to these target parameters was continually hand-tweaked according to human annotation of watched video clips captured along with the sensor data. This method of design is appropriate for this system as the goal is to label videos using sensor data similarly to how a human would label through viewing.

In general, the sensor data processing seeks to determine an activity level of the participant, indicating moments of change (physical, internal, situational) - it seeks to observe signaling in social situations that quantifies changes in the participants's relationships with others and with society, indicating important moments that need to be captured in video and included in any final product.

Examples of this derived data, taken from a live scenario,

where seven participants each wore a badge and a wrist sensor for one hour while within the coverage area of the SPINNER distributed camera system, are presented below in Figures 10-14. For more examples and larger format plots, please see [1].

5.4.1 Activity

The activity parameters indicate an action resulting in some kind of a change, be it a change in current task execution, mental state or attitude, and/or social situation. These changes present themselves in an observable and physical fashion and are therefore important labels as to how the audience will view the captured video clip.

The SPINNER wearable sensor devices have been designed to accurately detect motion with the inclusion of inertial sensors mounted on the wrist and on the torso. Each of these devices has a 3-axis accelerometer and a 2-axis gyroscope. The badge device also has a 3-axis compass reporting absolute angular orientation. Figure 10 shows the activity envelope calculated from 10 sensor data streams for one participant over a 3 minute interval.

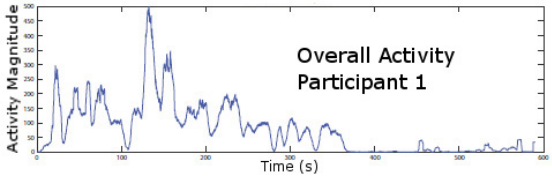

Figure 10: Subject's Overall Physical Activity Envelope

The wrist device is also equipped with a galvanic skin response sensor. This sensor works by measuring the skin conductivity to assess perspiration level. The moisture content in the skin fluctuates according to external temperature, physical activity, and all things being equal, emotional arousal. The GSR, as expected [29], tracks with the physical activity. In the case where the GSR rises above a threshold in a period of little physical activity, the SPINNER system will treat the internal activity as if it were physical activity, indicating some sort of situational change for the subject.

5.4.2 Social Signals

During a social interaction, we express significant amounts of information as to our situation, internal states, and personality. The SPINNER application creates a "relationship score" parameter ranging from a very negative relationship (i.e. active dislike) to a very positive relationship (i.e. active like), with the center point being no relationship (they have not met) or indifference. This relationship strength parameter is sometimes referred to as respect, social level, or level of attraction.

The relationship score is calculated from the level of interest/attention one person pays to another and the energy used when communicating. The interest level determines if the relationship value will be positive or negative. The interest level is then combined with the amount of energy expended in the conversation to determine the intensity of the relationship, be it negative or positive.

In order to calculate this value, several intermediate parameters need to be calculated for a social encounter, all of which can also be used to label the video clip. Many of these

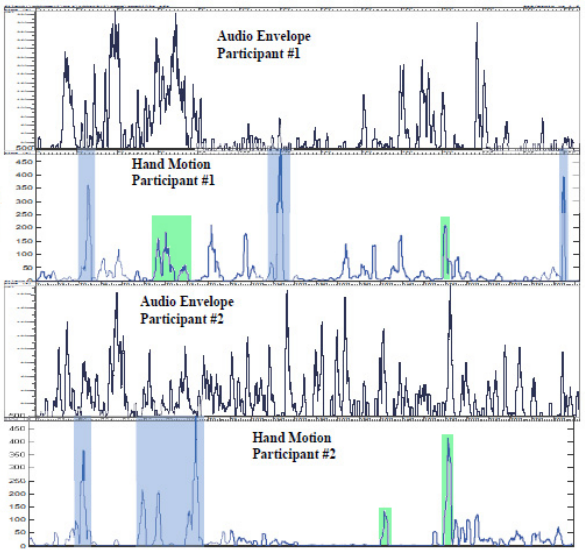

Figure 11: Hand activity aligned with audio envelope for two participants in a conversation, data taken from actual natural test situation. Blue shaded areas are time periods automatically selected when hand motion is high and speech envelope is low. These areas are considered as times of low attentiveness due to increased hand motion while listening. The green shaded areas on the plot are automatically created to show areas of emphatic gesture, when both the audio and hand activity are above thresholds showing moments of high speaker intensity

parameters have been determined from literature related to non-linguistic communication and human signaling [27] and then tested with the SPINNER system's video observation capabilities. To ascertain the level of interest, we take into consideration hand motion while listening, posture, body mirroring with speaker, body jitter while listening, and relative body orientation between speaker and listener. For communication energy, we mainly take into consideration hand motion while speaking to indicate emphatic speech, but we have also experimented with posture, body orientation, and conversation dynamics (interruptions, etc) to indicate speaking intensity. The raw data considered (audio envelope and physical activity from two particpants interacting) is shown in Figure 10 and Figures 11 through 14 illustrate the process to get from this raw data to an overall energy metric for the conversation.

The conversational energy of one participant talking to another is offset with values from the interest level of the other participant to determine the relationship score in each direction. Many of the fluctuations of the various envelopes used to calculate the relationship score are still present in the final values. This is exemplified by moments where no one is speaking or a lull between signals. Due to these fluctuations, the relationship score is interpolated to look at major trends, as opposed to using instantaneous data. The interpolation algorithm shown in blue in Figure 14 has been hand-tweaked based on observation of the videos from various social interactions.

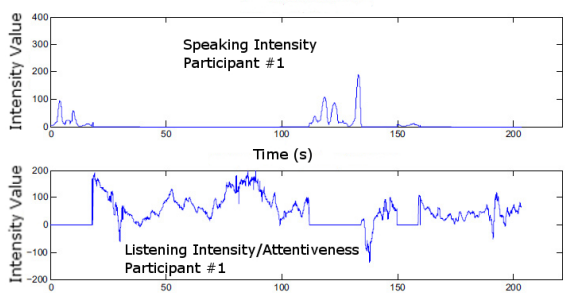

Figure 12: Participant 1's Speech Intensity and Listening Intensity/Attentiveness, the value is set to zero for speaking intensity while listening and set to zero for listening intensity while speaking

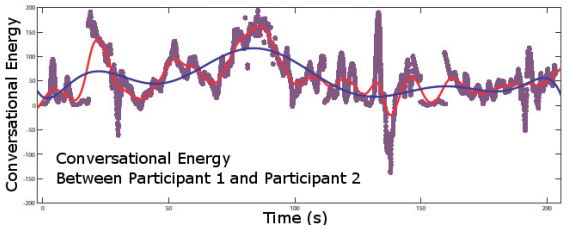

Figure 13: The conversation energy of Participant 1 is calculated by adding the zero-aligned speaking and listening energy values

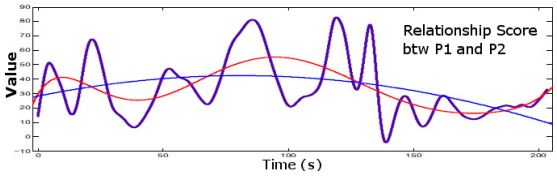

Figure 14: The relationship score over the course of one conversation between Participant 1 and Participant 2. In general the conversation goes downhill from there and dwindles into a fairly indifferent interaction

The activity and social derivations were evaluated by selecting some individual clips, posting them on YouTube, and asking viewers questions similar to what the sensors tell us. These results are shown in Figure 15.

Clip	SPINNER social	Survey social	SPINNER peak	Survey peak	Survey/ SPINNER peak time error	SPINNER relationship	Survey relationship
Clip 11	Social	91.7% social	N/A	N/A	N/A	-2	45
Clip 6	Not	60% not	81	74	1.3%	N/A	N/A
Clip 4	Not	100% not	50	53	5.6%	N/A	N/A
Clip 3	Social	90% social	N/A	N/A	N/A	45	84
Clip 9	Not	100% not	65	61	2.4%	N/A	N/A
Clip 12	Not	100% not	40	32	23.0%	N/A	N/A
Clip 13	Social	87.5% social	N/A	N/A	N/A	12	65

Figure 15: Results from the sensor processing evaluation comparing sensor derived values to human observed values, curves show tracking of results

5.5 User Input Parameters and Output Video Assembly

After the clips are captured, divided, and labeled, the SPINNER Video system can create an endless array of edited videos as determined by a set of input parameters. The first input parameters that control the creation of the final video from the labeled clips are the main character and the story trajectory. The story trajectory (see Figure 16 for example) is the overall shape of one or more of the parameters over the course of the final output video.

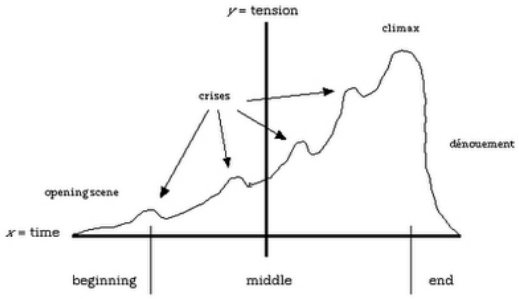

Figure 16: Example Story Trajectory, Y-Axis can be mapped to any parameter that the SPINNER system has available, according to the desired output video

In addition to one or more parametric story trajectories, the SPINNER system takes in the following parameters to control the desired video creation: curve matching tolerances for trajectories, ratio of social clips to non-social clips, whether reordering clips out of time order is allowed, and pacing parameters including target clip length, how to handle clips with little or no change, and which parameters are used to determine editing transition length and style.

The raw sensor data and derived information can also be mapped to the music mix, and specific edit points can be entered to help synchronize the audio and video.

6. EXECUTION, RESULTS, AND EVALUATION OF VIDEO CREATION WITH SENSOR NETWORKS

After months of repeated test runs and iterations of the platform and the sensor data processing methods, the system was executed stand-alone for three sessions ranging from one to five hours with a minimum of 40 operational portals. The first of these sessions used only the distributed network of cameras and sensors. The second and third sessions were focused on seven participants wearing badges and wrist sensor units.

In the video examples involving instrumented people, the system was set to automatically create an edited video based on one main character and driven by a set of parameters, such as an overall arc of social and non-social intensity. These videos were also posted to YouTube, and viewers were polled to assess their quality compared to one another and to a randomly created montage.

6.1 Created Video #1 - Activity

The first example video was generated without any wearable sensors to demonstrate the distributed video camera

system, the narrative trajectory matching, and final assembly of a video. The video was created by mapping the sensors from the Portals to the captured video. A clip's activity, or "story energy" level was calculated with the following formula:

$$activity = 0.50\frac{motion}{100} + 0.25\frac{sound}{32767} + 0.25badges \quad (1)$$

In the above equation, "motion" is defined as the percentage of time during the previous rolling window where the PIR sensor detected motion. "Sound" is defined as the amplitude of the sound pressure from the portal's microphone, and "badges" is either a one or a zero, depending on whether the portal has seen any badges recently.

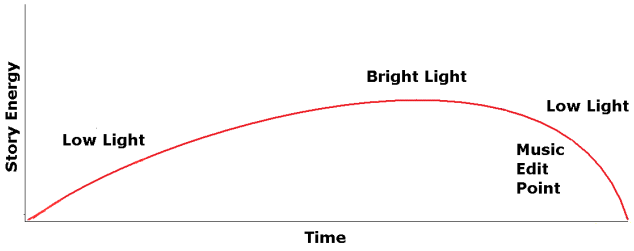

Figure 17: Energy/Activity Curve Entered for Activity Video Creation

In addition to the story energy trajectory for the video, the light level is mapped to the editing process. The clips selected in the first and last third of the final output are required to have lower light readings than the clips selected for the middle third. Taking into account the activity value and the light level, the clips are sequenced to best fit the trajectory shown in Figure 17.

Figure 18: Activity Video Still

The 1m30s video was created with thirty-five active cameras over the course of a large two-day, 300-person research symposium, resulting in over three weeks of raw video footage that the system needed to sort. Upon viewing this short video, those who were in attendance of the event said it accurately represented how they remember the event unfolding. Around 75% of those who were not in attendance said that the video made them wish they were there and provided insight into the feeling, dynamics, and structure of the event. A still from this video is shown in Figure 18 and the entire video is available for view on YouTube at [30].

6.2 Created Video #2 - Journey

This example video mainly illustrates the use of the wearable sensors to identify and locate the participants with respect to the cameras. This video was created from close to two hours of raw footage collected from thirty-five cameras with seven participants wearing sensors, one of which was arbitrarily selected as the main character.

The trajectory designed for this video was completely flat. In other words, this video was automatically assembled from all clips that have a steady energy curve. For this video, the energy level was mapped to inertial motion variance during non-social clips. The system was instructed to only select from non-social clips, except for one social clip in the middle. The social clip selected should be the one with the flattest energy curve. For the case of the social clips, the energy curve for this video was mapped to listening energy. The general idea of these particular settings is to test the systems ability to track a main character and create a story based around slowly searching and finding one social moment to interrupt an otherwise steady adventure.

The audio in this Journey Video is mapped to the motion of the main characterŠs wrist. The loudness of the background music in relation with the sync sound recorded from the subject's microphone increases with the amount of wrist activity. This type of mapping is useful to emphasize the subject's motion in the video through the use of sound and in general can heighten how engaging the video is for the audience.

Figure 19: Journey Video Still

The final journey video was around two minutes long edited down from over sixty hours of raw footage including one hour of raw footage that contained the main character. A still from this video is shown in Figure 19 and the entire video is available for view on YouTube at [31].

6.3 Created Video #3 - Story Arc

The next selected video example that was created by the SPINNER application illustrates the use of a standard story structure to sequence the clips. This system was asked to create a video characterized by a period of slow activity, followed by an up-tempo search, then ultimately leveling off, but at a higher level than the start, as would be expected from the experience of searching and finding.

The parameters used in creating this video are the wrist and body motion variance (while the main character is alone), and a combination of speaking energy from the main character and listening energy from other characters he is talking to

(while in a social interaction clip). The clips are sequenced according to this calculated energy parameter and whether or not the clip contained a social interaction to best fit the trajectory shown in Figure 20.

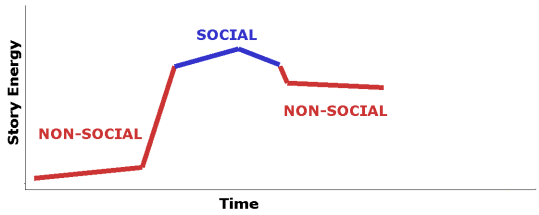

Figure 20: Story Arc Video Trajectory as Input to Video Creation Application

This video was created during the exact same time frame as the journey video from the previous section. Accordingly, it was also created from over sixty hours of raw footage with seven participants, including an hour of footage containing the main character. This video shows an entirely different perspective of events from the same time (and extracted from the same source footage) as the journey video. In each case, when the two main characters crossed paths, the specifics of their meeting did not match the programmed story conditions, hence this clip was not selected. This example illustrates how the wearable sensors can be used to project out media that represents a different series of events from a common shoot. A still from this video is shown in Figure 21 and the entire Story Arc Video can be viewed on YouTube at [32].

Figure 21: Story Arc Video Still

6.4 Evaluation of Video Creation

In order to evaluate the SPINNER Video Application, the videos that it has created have been uploaded to YouTube so that they may be viewed and rated by a large audience. The Journey Video and the Story Arc Video as described above have been uploaded along with a video made by a random assembly of clips also from the SPINNER cameras. The randomly assembled video is also available at [33].

Our results (Figure 22) show that both the Story Arc Movie and the Travel Movie were enjoyed substantially more than a randomly assembled video, despite all three movies coming from the same source material. Almost everyone selected the random movie as the one most likely assembled by a mechanized process. In general, the results show that the videos created with by the narrative-informed SPINNER system were enjoyed in general as well as in comparison to

Question	Travel Movie	Story Arc Movie	Random Crap
Enjoyment (1-10)	5.83	7.02	4.01
Memorable (1-10)	5.58	6.5	4.08
Insight (1-10)	5.58	6.42	3.08
Told a story (1-10)	5.18	6.09	3.09
Length (5 is perfect)	4.67	4.58	3.42
Mechanized feel (%)	0%	18.2%	82.8%

Figure 22: Results of User Survey

the video edited by a process without structure. One curious result was that the random video that people did not enjoy as much was seen being as too short. This is probably due to the fact that it has no meaning, hence viewers assumed that something was missing that would be included if the video was longer. The other two videos were considered as near perfect length, indicating that they were viewed as being a complete story with nothing missing. These results, combined with the fact that it would take a human hundreds of hours to filter through the video and recreate this process, indicate that these techniques are promising and the SPINNER Video Application was successful.

7. CONCLUSIONS

Media technology has exploded into the very fabric of our lives, bringing with it a host of potential benefits and drawbacks. Through the design and implementation of our system and our experimentation with automatic user-generated video content creation we have started to understand and confront the drawbacks inherent with the coming age of truly ubiquitous media and allow its benefits to have their full impact on society.

More specifically, we have developed and deployed a powerful new media platform for experimentation with distributed cameras, wearable sensing, social signal analysis, situated media technologies such as ubiquitous displays, wearable distributed audio capture, and the use of narratology to design applications and systems. Using this platform, we have created a new form of entertainment and communication allowing users to create comprehensible video content out of their social and personal behavior. This method of creation promotes validity of output and puts the ability to create meaningful content into everyone's hands - in other words, no special knowledge or practice is required to create it, as the result arises from one's actual experience. Our system has successfully transformed the environment into a creative tool and a new technique for self-documentation and self-reflection that can lead to new and unexpected insights about events that might go unnoticed while living them.

A major point of future work is to execute the system for an extended period of time on the order of weeks providing an extended set of data to label, filter, and ultimately have better options for which captured events to include in the final video.

While we have demonstrated a complete content creation system, one can do much more with these tools given enough time to really explore some deeper areas, such as using the wearable sensors to reliably label video with true emotional parameters (i.e. conflict and respect). We have started to work on these types of mappings as well as using sensor

data to identify moments that can be considered as specific story beats according to a particular model of narrative. The system could be greatly improved by adding the recognition of specific gestures and relating these to story beats. These enhancements will allow future versions of the SPINNER video content creation system to create more interesting and efficient output. However, even in its simplest form, the demonstrated system as-is would be an extremely useful editor/producer's aid, automatically screening out irrelevant footage and suggesting clips that could fit a storyboard.

8. REFERENCES

[1] Laibowitz, M. Creating Cohesive Video with the Narrative-Informed use of Ubiquitous Wearable and Imaging Sensor Networks. PhD Thesis. January 2010.

[2] N. W. Gong, M. Laibowitz and J. A. Paradiso. Experiences and Challenges in Deploying Potentially Invasive Sensor Systems for Dynamic Media Applications. Cloud-Mobile Convergence for Virtual Reality Workshop. 2010.

[3] K. S. J. Pister, J. M. Kahn and B. E. Boser. Smart dust: communicating with a cubic-millimeter computer. Computer, 34(1):44-51, Jan. 2001.

[4] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, Ken Wood. SenseCam: a Retrospective Memory Aid In Dourish and A. Friday (Eds.): Ubicomp 2006, LNCS 4206, pp.177-193, 2006.

[5] http://www.vicon.com/SenseCam

[6] Wolf, Ozer, Lv. Smart cameras for embedded systems. In IEEE Computer, volume 35, pages 48-53, September 2002

[7] Thiago Teixeira, Dimitrios Lymberopoulos, Eugenio Culurciello, Yiannis Aloimonos, Andreas Savvides. A Lightweight Camera Sensor Network Operating on Symbolic Information, Proceedings of First Workshop on Distributed Smart Cameras 2006, ACM SenSys 2006

[8] Lymberopoulos, D, Teixeira, T, Savvides, A. Macroscopic Human Behavior Interpretations Using Distributed Imagers and Other Sensors, under review for publication in Transactions of IEEE

[9] Wu-chi Feng, Brian Code, Ed Kaiser, Mike Shea, Wu-chang Feng. Panoptes: scalable low-power video sensor networking technologies. 11th ACM Multimedia, pages 90-91, New York, NY, USA, 2003.

[10] Gordon Bell, Jim Gemmell. "A Digital Life," Scientific American, 296(3), February 2007, pp. 58-65.

[11] http://erc.europa.eu/

[12] T. Choudhury and A. Pentland. Sensing and modeling human networks using the sociometer. In Wearable Computers, 2003. pages 216-222, Oct. 2005.

[13] Eagle, N and Pentland, A. Reality mining: sensing complex social systems. Personal Ubiquitous Comput., 10(4):255-268, 2006.

[14] Mathew Laibowitz, Jonathan Gips, Ryan Aylward, Alex Pentland, Joseph A. Paradiso. A sensor network for social dynamics. IPSN 2006. pp. 483-491, 19-21 April 2006.

[15] Daniel Olguin Olguin, Benjamin N. Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex (Sandy) Pentland. Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior. IEEE Trans. on Systems, Man, and Cybernetics - Part B, 39(1), Feb. 2009, pp. 43-55.

[16] Singh and Williams. Lifenet: a propositional model of ordinary human activity. In Proceedings at K-CAP 2003, 2003.

[17] Weld et al. Automatically personalizing user interfaces. In IJCAI03, Acapulco, Mexico, August 2003.

[18] Guensler R. Wolf J and Bachman W. Elimination of the travel diary: an experiment to derive trip purpose from gps travel data. In Transportation Research, 2001.

[19] Beth Logan, Jennifer Healey, Matthai Philipose, Emmanuel Tapia, and Stephen Intille. "A long-term evaluation of sensing modalities for activity recognition," in Proceedings of the International Conference on Ubiquitous Computing, vol. LNCS 4717. Springer, 2007, pp. 483Ű500.

[20] David Bannach, Oliver Amft, and Paul Lukowicz, "Rapid Prototyping of Activity Recognition Applications," IEEE Pervasive Computing, pp. 22-31, April, 2008

[21] Duda et al. Pattern Classification (2nd Edition). Wiley-Interscience, November 2000.

[22] E. Wertheim. Historical background of organizational behavior.

[23] V. M. Bove and J. Mallett. Collaborative knowledge building by smart sensors. BT Technology Journal, 22(4):45-51, 2004.

[24] J. Mallett and V. M. Bove. Eye society. In Multimedia and Expo, 2003. ICME Š03. pages 6-9 July 2003.

[25] Reddy, S., Parker, A., Hyman, J., Burke, J., Estrin, D., and Hansen, M. "Image browsing, processing, and clustering for participatory sensing: lessons from a DietSense prototype," Proc. of EmNets 2007, Cork Ireland, pp. 13-17.

[26] Sayed Ahmad; Rasit Eskicioglu; Peter Graham; "Design and Implementation of a Sensor Network Based Location Determination Service for use in Home Networks," Mobile Adhoc and Sensor Systems (MASS), 2006 pp.622-626, Oct. 2006

[27] Oksana Bulgakowa dir., The Factory of Gestures - Body Language in Film. PPMedia and Stanford Humanities Lab. DVD. 160 minutes. 2007.

[28] Laibowitz and Paradiso. "The UbER-Badge, A Versatile Platform at the Juncture Between Wearable and Social Computing," in Advances in Pervasive Computing, OCG, 2004, pp.363-368.

[29] Picard et al. The handwave bluetooth skin conductance sensor. In ACII, volume 3784 of Lecture Notes in Computer Science, pages 699.706. Springer, 2005.

[30] http://www.youtube.com/watch?v=rDv7D2xeKrs

[31] http://www.youtube.com/watch?v=vw9outkeR1k

[32] http://www.youtube.com/watch?v=ME1a0gSvMD8

[33] http://www.youtube.com/watch?v=Wa4dXm7_yAE