(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 June 2003 (12.06.2003)

PCT

(10) International Publication Number
WO 03/049430 A2

(51) International Patent Classification[7]: H04N 5/76

(21) International Application Number: PCT/IB02/04944

(22) International Filing Date:
20 November 2002 (20.11.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/011,872 6 December 2001 (06.12.2001) US

(71) Applicant: **KONINKLIJKE PHILIPS ELECTRONICS N.V.** [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).

(72) Inventors: **DIMITROVA, Nevenka**; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). **ZIMMERMAN, John**; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). **MCGEE, Thomas**; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). **JASINSCHI, Radu, S.**; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).

(74) Agent: **GROENENDAAL, Antonius, W., M.**; Internationaal Octrooibureau B.V., Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).

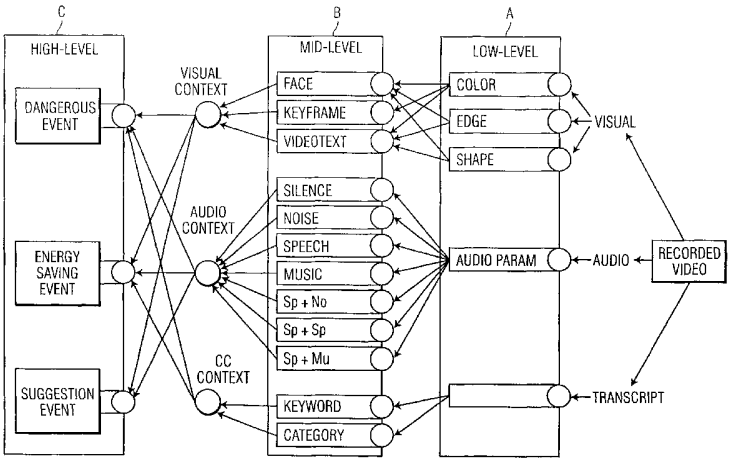
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: ADAPTIVE ENVIRONMENT SYSTEM AND METHOD OF PROVIDING AN ADAPTIVE ENVIRONMENT



(57) Abstract: An adaptive environment system comprises a recording device for recording a video which is analyzed by a processor and indexed according to features of the video. The video is segmented into at least visual, audio, and textual components, which can be analyzed by the processor. The processor then creates an index file of the features analyzed and stores the video along with the index file on a storage device. The video can then be searched according to the index file and a portion of the video identified by the search returned to a display for viewing. In addition, the adaptive environment system may comprise a processing system connectable to a network wherein the network comprises one or more interconnected sensors. The processing system comprises a computer readable medium comprising computer code for instructing one or more processors to: (a) receive recorded data from the one or more sensors connectable to the processing system; (b) analyze the recorded data to identify an event occurring in the recorded data; (c) determine whether a response to the identified event is appropriate; and (d) when a response is appropriate generate a signal associated with the response.

WO 03/049430 A2

Adaptive environment system and method of providing an adaptive environment

The present invention relates to the a system for providing an adaptive environment and, in particular, to a system for use in an environment to record, segment, and index video, audio, and other data captured by sensors in the environment.

As analog and digital recording of both audio and video have become mainstream, people are increasingly recording various events in their lives. Video/audio tapes, and more recently CDROMs, are a cumbersome means of storing and cataloging events. Oftentimes, tapes are lost or the label describing the contents becomes unreadable. Even when a tape is found, the user often has to fast forward through hours of video before finding the desired event. While it may be easier to store and identify individual files in digital form, generally available indexing systems are limited and do not adequately provide for the segmentation and indexing of events on a frame-by-frame basis.

Other systems for recording and indexing television programs, such as personal video recorders (PVRs) like TiVo®, use electronic program guide metadata to automatically select and store whole TV programs based on users' profiles. These systems can be limited, however, because such systems do not allow for the segmentation and indexing of events on a frame-by-frame basis.

Furthermore, events that take place in one's house or office may be missed (i.e., unrecorded) because there are no tapes or the camera is out of battery. For example, a child's first words or first steps could be missed because by the time the camera is ready the event has passed.

Home security and home monitoring systems are also known. Such systems use motion detectors, microphones, cameras, or other electronic sensors to detect the presence of someone when the system is armed. Other types of home monitoring systems employ a variety of sensors to monitor various home appliances, including furnaces, air conditioners, refrigerators, and the like. Such systems, however, are generally limited in their use due to the specialized nature of the sensors and the low processing power of the

processors powering such systems. For instance, home alarms are routinely falsely set-off when a household member or the family dog strays into the sight of a motion detector.

Furthermore, current systems for denying access to certain home appliances, such as the television or personal computers attached to the Internet, are cumbersome and ineffective. For instance, some televisions can be programmed to require a password to access television programs of a certain rating. These systems, however, require that each member of a family use a PIN to identify themselves to the television. Oftentimes, such systems go unused because people find it unmanageable to use such systems.

Thus, a system that passively records data and provides for the segmentation and indexing of the data such that it is easily retrievable is desirable.

It is further desirable to have a home or office security system that could identify individuals and avoid false alarms. Moreover, it is desirable to use such a system to control access to household appliances such as televisions, Internet connections, personal computers, ovens, etc.

Yet further, it is desirable to provide for a system that can observe the behavior and habits of individuals and anticipate their actions. For example, a system that could control repetitive tasks such as controlling the heating, cooling, lighting, and other household and office conditions based upon an individual's preferences or past behavior is desirable.

The present invention overcomes shortcomings found in the prior art. The present invention provides an integrated and passive adaptive environment that analyzes audio, visual, and other recorded data to identify various events and determine whether an action needs to be taken in response to the event. The analysis process generally comprises monitoring of an environment, segmentation of recorded data, identification of events, and indexing of the recorded data for archival purposes.

Generally speaking, one or more sensors monitor an environment and passively record the actions of subjects in the environment. The sensors are interconnected with a processing system via a network. The processing system is advantageously operative with a probabilistic engine to segment the recorded data. The segmented data can then be analyzed by the probabilistic engine to identify events and indexed and stored in a storage device, which is either integrated with or separated from the processing system. As will become evident from the following disclosure, the processing system according to the present

invention can perform any number of functions using the probabilistic approach described herein.

In one embodiment of the invention, the processing system segments and indexes the recorded data so as to allow users to search for and request events that have occurred in the environment. For example, users can request particular events that have occurred in the operating environment, which are extracted from the stored data and replayed for the user. In addition, the system of the present invention monitors the repeated behavior of subjects in the environment to learn their habits. In a further embodiment of the present invention, the system can remind the subject to perform a task or even perform that task for the subject.

The processing system is connectable to a network of sensors, which passively record events occurring in the environment. In an embodiment of the present invention, the sensors or recording devices may be video cameras capable of capturing both video and audio data or microphones. Preferably, the sensors are connected to a constant source of power in the operating environment so as to passively operate on a consistent basis. As the data is captured, it separates the video from the audio data captured by the cameras. These separated streams of data are then analyzed by a probabilistic engine of the processing system, which analyzes the streams of data to determine the proper segmentation and indexing of the data.

The probabilistic engine of the processing system also enables the processing system to track repetitive actions by the recorded subjects. The probabilistic engine can then select those activities that occur more frequently than other of the subject's activities. Thus, the probabilistic engine essentially learns the habits of the subjects it records and can begin to remind the subjects to perform tasks or perform tasks automatically.

In another embodiment, the system operates as a security system wherein the processing system uses the captured data to identify individuals and provide or deny access to various components of the operating environment. Once an individual is identified, the processing system can access a database of stored user access parameters. For instance, a young child may not be provided access to certain channels on the television. Thus the processing system can automatically identify the young child and set a system in the television (such as a V-chip) to deny access to certain channels based upon this user information. Furthermore, the system can identify when unidentified individuals are present in the house and notify the proper authorities or set off an alarm.

In accordance with another aspect of the present invention, a method of retrieving recorded events is provided, where the method comprises collecting data from various recording devices, de-mixing the data into individual components, analyzing each component of the de-mixed data, segmenting the analyzed data into a plurality of components, indexing the segmented data according to a set of values collected by the processing system, and retrieving the data from a storage device in response to a request from a user that includes an identifier of a portion of the indexed and segmented data.

The above and other features and advantages of the present invention will become readily apparent from the following detailed description thereof, which is to be read in connection with the accompanying drawings.

In the drawing figures, which are merely illustrative, and wherein like reference numerals depict like elements throughout the several views:

Fig. 1 is a schematic diagram of an overview of an exemplary embodiment of the system architecture in accordance with the present invention;

Fig. 2 is a flow diagram of an exemplary process of segmenting and classifying recorded data;

Fig. 3 is a schematic diagram of an exemplary embodiment of the segmentation of the video, audio, and transcript streams;

Fig. 4 is a flow diagram of an exemplary process of creating an index file for searching recorded data;

Fig. 5 is a schematic diagram of an exemplary process of retrieving indexed data; and

Fig. 6 is a flow diagram of an exemplary process of providing security to electronic devices connected to the system of the present invention.

The present invention provides an adaptive environment that comprises a passive event recording system that passively records events occurring in the environment, such as a house or office. The recording system uses one or more recording devices, such as video cameras or microphones. The system processes the recorded events to segment and index the events according to a set of parameters. Because the system is passive, people interacting with the system need not concern themselves with the operation of the system.

Once the recorded data is segmented and indexed it is stored on a storage device so as to be easily retrievable by a user of the system.

The passive recording system according to the present invention preferably comprises one or more recording devices for capturing a data input and a processing engine, also referred to as a processing system or a processor, communicatively connected to the recording devices. Once content is received from the recording devices, the processing engine segments the content according to a three-layered approach that uses various components of the content. The segmented content is then classified based on the various content components. The content is then stored on a storage device that is also interconnected to the processor via a network such as a local area network (LAN). The content can be retrieved by users by searching for objects that are identifiable in the content, such as searching for a "birthday and Steve". In such an example, the processing engine would search for segments of the content fulfilling the search criteria. Once found the entire segments can be returned to the user for viewing.

The processing system preferably uses a Bayesian engine to analyze the data stream inputs. For example, preferably each frame of the video data is analyzed so as to allow for the segmentation of the video data. Such methods of video segmentation include but are not limited to cut detection, face detection, text detection, motion estimation/segmentation/detection, camera motion, and the like. Furthermore, the audio data is also analyzed. For example, audio segmentation includes but is not limited to speech to text conversion, audio effects and event detection, speaker identification, program identification, music classification, and dialogue detection based on speaker identification. Generally speaking, audio segmentation involves using low level audio features such as bandwidth, energy and pitch of the audio data input. The audio data input may then be further separated into various components, such as music and speech. Using these and other parameters, the system passively records and identifies various events that occur in the home or office and can index the events using information collected from the above processes. In this way, a user can easily retrieve individual events and sub events using plain language commands or the processing system can determine whether an action is necessary in response to the identified event. In operation, upon receipt of a retrieval request from a user, the processing engine calculates a probability of the occurrence of an event based upon the plain language commands and returns the requested event.

By way of example, as shown in Fig. 3, the probabilistic engine can identify dangerous events (such as burglaries, fires, injuries, etc.), energy saving events (such as

opportunities to shut lights and other appliances off, lower the heat, etc.), and suggestion events (such as locking the doors at night or when people leave the environment).

It should be understood that although the present invention is described in connection with the passive recording system used in an operating environment, such as a home or office type environment, the passive recording system can be used in any operating environment in which a user wishes to record and index events occurring in that environment. That environment may be outdoors or indoors.

Refer now to Fig. 1, a system 10 according to the present invention is shown wired in a household environment 50. As can be seen, the house has many rooms 52 that may each have a separate recording device 12. Each recording device 12 is interconnected via a local area network (LAN) 14 to one another and to the processor 16. In turn, the processor 16 is interconnected to a storage device 18 for storing the collected data. A terminal for interacting with the processor 16 of the passive recording system 10 may also be present. In a preferred embodiment, each recording device 12 is wired to the house's power supply (not shown) so as to operate passively without interaction from the users. Thus, the recording system 10 operates passively to continuously record events occurring in the house without intervention or hassle by the users. Furthermore, one or more of the electronic systems (not shown) in the operating environment (e.g., appliances, televisions, heating and cooling units, etc.) may be interconnected to the LAN 14 so as to be controllable by the processor 16.

The processor 16 is preferably hosted in a computer system that can be programmed to perform the functions described herein. By way of example only, the computer system may comprise a control processor and associated operating memory (RAM and ROM), and a media processor, such as the Philips TriMedia™ Tricodec card for pre-processing the video, audio and text components of the data input. The processor 16, which may be, for example, an Intel Pentium chip or other multiprocessor, performs an analysis of frames of data captured by the recording devices to build and store an index in an index memory, such as, for example, a hard disk, file, tape, DVD, or other storage medium. The computer system is interconnected to and communicates with the storage device 18, recording devices 12, and other electronic components via a LAN 14, which is either hardwired throughout the operating environment or operates wirelessly.

Operatively coupled to the processor 16 is the storage device 18 (for example, RAM, hard disk recorder, optical storage device, or DVHS, each preferably having hundreds of giga-bytes of storage capability) for storing the recordings of the events. Of course, the processor 16 and storage device 18 can be integrated into a single unit.

The recording devices or sensors 12 may be video cameras having integrated microphones so as to receive both video and audio data. In other embodiments the recording devices 12 may be microphones, motion detectors or other types of sensors. The recording devices 12 may further be equipped to have motion detectors that enable the recording device 12 to fall into a sleep mode when no events are occurring in a particular room and awake upon the detection of movement or action in the room. In this way, power will be conserved and storage space in the storage device 18 is preserved. Yet further, the video cameras can include a pivoting system that allows the cameras to track events occurring in a particular room. In such a system, by way of example, a child that is walking from a bedroom can be followed out the door by a first camera, down the hallway by a second camera and into a play area by a third camera.

Each camera would pivot to follow the child's movements and then turn off when the movement ceased to occur for a preset time period in that particular room. The now active camera would then detect the motion of the child entering the area and begin recording. This tracking feature of the recording devices 12 will be further described below in connection with an embodiment of the invention involving a content distribution system.

An exemplary method of tracking a subject in a multiple camera system is described in International Publication WO 00/08856 to Sengupta et al., of such a camera tracking system generally comprises two or more video cameras 12 (shown in Figure 1). The cameras 12 may be adjustable, pan/tilt/zoom, cameras. The cameras 12 provide an input to a camera handoff system (not shown in the Figures); the connections between the cameras 12 and the camera handoff system may be direct or remote, for example, via a telephone connection or other network. The camera handoff system preferably includes a controller, a location determinator, and a field of view determinator. The controller effects the control of the cameras 12 based on inputs from various sensors, the location determinator and the field of view determinator.

The environment 50 also preferably includes an integrated speaker or monitor system 30 interconnected with the LAN 14. As will be described further below, the monitor/speaker system 30 can be used to broadcast content to users of the system 10, such as TV, video, audio, or even voice reminders.

With reference to Fig. 2, an overview of the process of capturing, analyzing, segmenting, and archiving the content for retrieval by the user is shown. When the recording devices are activated, video content is captured by the recording devices and transmitted to the processor, in steps 202 and 204. The processor receives the video content as it is

transmitted and de-multiplexes the video signal to separate the signal into its video and audio components, in step 206. Various features are then extracted from the video and audio streams by the processor, in step 208.

As shown in Fig. 3, the features of the video and audio streams are preferably extracted and organized into three consecutive layers: low A, mid B and high C level. Each layer has nodes with associated probabilities. Arrows between the nodes indicate a causal relationship. The low-level layer A generally describes signal-processing parameters. In an exemplary embodiment the parameters include but are not limited to: the visual features, such as color, edge, and shape; audio parameters such as average energy, bandwidth, pitch, mel-frequency cepstral coefficients, linear prediction coding coefficients, and zero-crossings. The processor then preferably combines the low-level features to create the mid-level features. The mid-level features B are preferably associated with whole frames or collections of frames while low-level features A are associated with pixels or short time intervals. Keyframes (first frame of a shot, or a frame that is judged important), faces, and videotext are examples of mid-level visual features; silence, noise, speech, music, speech plus noise, speech plus speech, and speech plus music are examples of mid-level audio features; and keywords of a transcript along with associated categories make up the mid-level transcript features. High-level features C describe semantic video content obtained through the integration of mid-level features across the different domains. In other words, the high level features represent the classification of segments according to user or manufacturer defined profiles, described further below.

With reference again to Figure 2, the processor attempts to detect whether the audio stream contains speech, in step 210. An exemplary method of detecting speech in the audio stream is described below. If speech is detected, then the processor converts the speech to text to create a time-stamped transcript of the recorded content, in step 212. The processor then adds the text transcript as an additional stream to be analyzed (see Fig. 3), in step 214.

Whether speech is detected or not the processor then attempts to determine segment boundaries, i.e., the beginning or end of a classifiable event, in step 216. In a preferred embodiment, the processor performs significant scene change detection first by extracting a new keyframe when it detects a significant difference between sequential I-frames of a group of pictures. The frame grabbing and keyframe extracting can also be performed at pre-determined intervals. The video pre-processing module of the processing engine employs a DCT based implementation for frame differencing using cumulative macroblock difference measure. Alternatively, a histogram based method may be employed.

We should note here that video material from home video cameras and surveillance cameras is quite different from broadcast video and some of the methods for keyframe extraction applied on broadcast video would not be effective in the home area. However, any method that can detect a significant difference between subsequent frames and help in extraction of important frames can be employed in the system. Unicolor keyframes or frames that appear similar to previously extracted keyframes get filtered out using a one-byte frame signature. The processing engine bases this probability on the relative amount above the threshold using the differences between the sequential I-frames.

A method of frame filtering is described in U.S. Patent No. 6,125,229 to Dimitrova et al. and is briefly described below. Generally speaking the processor receives content and formats the video signals into frames representing pixel data (frame grabbing). It should be noted that the process of grabbing and analyzing frames is preferably performed at pre-defined intervals for each recording device. For instance, when a recording device begins recording data, keyframes can be grabbed every 30 seconds. In this way, the processing engine can perform a Bayesian probability analysis, described further below, to categorize an event and create an index of the recorded data.

Once these frames are grabbed every selected keyframe is analyzed. Video segmentation is known in the art and is generally explained in the publications entitled, N. Dimitrova, T. McGee, L. Agnihotri, S. Dagtas, and R. Jasinschi, "On Selective Video Content Analysis and Filtering," presented at SPIE Conference on Image and Video Databases, San Jose, 2000; and "Text, Speech, and Vision For Video Segmentation: The Infomedia Project" by A. Hauptmann and M. Smith, AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision 1995. Any segment of the video portion of the recorded data including visual (e.g., a face) and/or text information relating to a person captured by the recording devices will indicate that the data relates to that particular individual and, thus, may be indexed according to such segments. As known in the art, video segmentation includes, but is not limited to:

Significant scene change detection: wherein consecutive video frames are compared to identify abrupt scene changes (hard cuts) or soft transitions (dissolve, fade-in and fade-out). An explanation of significant scene change detection is provided in the publication by N. Dimitrova, T. McGee, H. Elenbaas, entitled "Video Keyframe Extraction and Filtering: A Keyframe is Not a Keyframe to Everyone", Proc. ACM Conf. on Knowledge and Information Management, pp. 113-120, 1997.

Face detection: wherein regions of each of the video frames are identified which contain skin-tone and which correspond to oval-like shapes. In the preferred embodiment, once a face image is identified, the image is compared to a database of known facial images stored in the memory to determine whether the facial image shown in the video frame corresponds to the user's viewing preference. An explanation of face detection is provided in the publication by Gang Wei and Ishwar K. Sethi, entitled "Face Detection for Image Annotation", Pattern Recognition Letters, Vol. 20, No. 11, November 1999.

Motion Estimation/Segmentation/Detection: wherein moving objects are determined in video sequences and the trajectory of the moving object is analyzed. In order to determine the movement of objects in video sequences, known operations such as optical flow estimation, motion compensation and motion segmentation are preferably employed. An explanation of motion estimation/segmentation/detection is provided in the publication by Patrick Bouthemy and Francois Edouard, entitled "Motion Segmentation and Qualitative Dynamic Scene Analysis from an Image Sequence", International Journal of Computer Vision, Vol. 10, No. 2, pp. 157-182, April 1993.

The method also includes segmentation of the audio portion of the video signal wherein the audio portion of the video is monitored for the occurrence of words/sounds that are relevant to the viewing preferences. Audio segmentation includes the following types of analysis of video programs: speech-to-text conversion, audio effects and event detection, speaker identification, program identification, music classification, and dialog detection based on speaker identification.

Audio segmentation includes division of the audio signal into speech and non-speech portions. The first step in audio segmentation involves segment classification using low-level audio features such as bandwidth, energy and pitch. Channel separation is employed to separate simultaneously occurring audio components from each other (such as music and speech) such that each can be independently analyzed. Thereafter, the audio portion of the video (or audio) input is processed in different ways such as speech-to-text conversion, audio effects and events detection, and speaker identification. Audio segmentation is known in the art and is generally explained in the publication by E. Wold and T. Blum entitled "Content-Based Classification, Search, and Retrieval of Audio", IEEE Multimedia, pp. 27-36, Fall 1996.

Speech-to-text conversion (known in the art, see for example, the publication by P. Beyerlein, X. Aubert, R. Haeb-Umbach, D. Klakow, M. Ulrich, A. Wendemuth and P. Wilcox, entitled "Automaic Transcription of English Broadcast News", DARPA Broadcast

News Transcription and Understanding Workshop, VA, Feb. 8-11, 1998 can be employed once the speech segments of the audio portion of the video signal are identified or isolated from background noise or music. The speech-to-text conversion can be used for applications such as keyword spotting with respect to event retrieval.

Audio effects can be used for detecting events (known in the art, see for example the publication by T. Blum, D. Keislar, J. Wheaton, and E. Wold, entitled "Audio Databases with Content-Based Retrieval", Intelligent Multimedia Information Retrieval, AAAI Press, Menlo Park, California, pp. 113-135, 1997. Events can be detected by identifying the sounds that may be associated with specific events. For example, the singing of "Happy Birthday" could be detected and the segment could then be indexed as a birthday event.

Speaker identification (known in the art, see for example, the publication by Nilesh V. Patel and Ishwar K. Sethi, entitled "Video Classification Using Speaker Identification", IS&T SPIE Proceedings: Storage and Retrieval for Image and Video Databases V, pp. 218-225, San Jose, CA, February 1997 involves analyzing the voice signature of speech present in the audio signal to determine the identity of the person speaking. Speaker identification can be used, for example, to search for a particular family member.

Event identification involves analyzing the audio portion of the data signal captured by the recording devices to identify and classify an event. This is especially useful in cataloging and indexing of events. The analyzed audio portion is compared to a library of event characteristics to determine if the event coincides with known characteristics for a particular event.

Music classification involves analyzing the non-speech portion of the audio signal to determine the type of music (classical, rock, jazz, etc.) present. This is accomplished by analyzing, for example, the frequency, pitch, timbre, sound and melody of the non-speech portion of the audio signal and comparing the results of the analysis with known characteristics of specific types of music. Music classification is known in the art and explained generally in the publication entitled "Towards Music Understanding Without Separation: Segmenting Music With Correlogram Comodulation" by Eric D. Scheirer, 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY October 17-20, 1999.

The various components of the video, audio, and transcript text are then analyzed according to a high level table of known cues for various event types, in step 218.

Each category of event preferably has a knowledge tree that is an association table of keywords and categories. These cues may be set by the user in a user profile or pre-determined by a manufacturer. For instance, the "graduation" tree might include keywords such as school, graduation, cap, gown, etc. In another example, a "birthday" event can be associated with visual segments, such as birthday candles, many faces, audio segments, such as the song "Happy Birthday", and text segments, such as the word "birthday". After a statistical processing, which is described below in further detail, the processor performs categorization using category vote histograms. By way of example, if a word in the text file matches a knowledge base keyword, then the corresponding category gets a vote. The probability, for each category, is given by the ratio between the total number of votes per keyword and the total number of votes for a text segment.

In a preferred embodiment, the various components of the segmented audio, video, and text segments are integrated to index an event. Integration of the segmented audio, video, and text signals is preferred for complex indexing. For example, if the user desires to retrieve a speech given during someone's birthday, not only is face recognition required (to identify the actor) but also speaker identification (to ensure the actor on the screen is speaking), speech to text conversion (to ensure the actor speaks the appropriate words) and motion estimation-segmentation-detection (to recognize the specified movements of the actor). Thus, an integrated approach to indexing is preferred and yields better results.

In step 220, this segment information is then stored along with the video content on a storage device connected to the processor.

A preferred process of generating the high level inferences of the high-level layer will now be described. Preferably, a Bayesian probabilistic analysis approach is used because such approach integrates either intra or inter-modalities. Intra-modality integration refers to integration of features within a single domain. For example: integration of color, edge, and shape information for videotext represents intra-modality integration because it all takes place in the visual domain. Integration of mid-level audio categories with the visual categories face and videotext offers an example of inter-modalities because it combines both visual and audio information to make inferences about the content. A probabilistic approach to this integration is found in Bayesian networks. They allow the combination of hierarchical information across multiple domains and handle uncertainty. Bayesian networks are directed acyclical graphs (DAG) in which the nodes correspond to (stochastic) variables. The arcs describe a direct causal relationship between the linked variables. The strength of these links is given by conditional probability distributions (cpds). More formally, let the set $\Omega(x_1, ...x_N)$

of N variables define a DAG. For each variable there exists a sub-set of variables of Ω, Π_{xi}, the parents set of x_i, i.e., the predecessors of x_i in the DAG, such that $P(x_i \mid \Pi_{xi}) = P(x_i \mid x_{i,...,} x_{i-1})$, where $P(\bullet \mid \bullet)$ is a cpd, strictly positive. Now, given the joint probability density function (pdf) $P(x_{i,...,} x_N)$, using the chain rule, we get that $P(x_{i,...,} x_N) = P(x_N \mid x_{N-1,...,} x_1)* ... * P(x_2 \mid x_1)P(x_1)$. According to this equation, the parent set Π_{xi} has the property that x_i and $\{x_1, ..., x_N\}\backslash\Pi_{xi}$ are conditionally independent given Π_{xi}.

As previously described the structure of a DAG is preferably made up of three layers. In each layer, each element corresponds to a node in the DAG. The directed arcs join one node in a given layer with one or more nodes of the preceding layer. Two sets of arcs join the elements of the three layers. For a given layer and for a given element, a joint pdf is calculated as previously described. There can exist an overlap between the different parent sets for each level.

Topic segmentation and classification performed by the processor as shown in the third layer (high-level C) of Fig. 3. In a preferred embodiment, the processor performs indexing of content according to the users' or a manufacturer's predefined high-level keyword table. The processor indexes the content by (i) reading keywords and other data from the high-level table and (ii) classifying the content into segments based on several high-level categories.

Thus, with reference to Fig. 4, there is shown an exemplary analysis of a conversation between two members of a household according to the present invention. Once the content is segmented and analyzed according a preferred embodiment, described above, a Bayesian approach or other probabilistic analysis approach may be used to create an index file for the segmented content. As can be seen, one method of indexing the event takes into account the appearance of visual, audio, and textual indicia of a particular event.

In this analysis, the processor determines the probability that an event fits into a category, which, as described above, includes a number of indicia of that category. The processor may additionally identify those subjects appearing in the visual segments using a face detection method. This information is stored in the index file and provides a link to the segmented content, which can be searched by a user.

By way of example only, with reference to Fig. 4, a conversation in the kitchen involving Bob and Mary regarding a certain stock "XYZ Corp." can be indexed as follows. In steps 402 and 404, the processor, after analyzing the various video, audio, and textual components, would record certain static data about the event. For instance, the date and time of the event and the room in which the event was captured would be stored in a

index file. Furthermore, the processor preferably uses a combination of the face detection segment of the video stream, along with a voice recognition segment of the audio stream to identify the subjects (Bob and Mary) associated with the event, in step 406. In steps 408 and 410, the processor would also categorize the event according to the textual terms that were repeated more than a certain number of times during the event. For example, an analysis of the text transcript would identify that the terms "XYZ Corp.", "stock", and "money" were repeatedly spoken by the subjects and, thus would be added to the index file. Moreover, the processor would use a probabilistic approach to determine the nature of the event, i.e., a conversation, in step 412. This is preferably performed by using predefined indicia of a conversation, including but not limited to the noise level and speech characteristics of the audio stream, the repeated changing of speakers in the text stream, and the limited movement of the subjects in the video stream.

With further reference to Fig. 5, an exemplary process of retrieving Bob and Mary's conversation is shown. As noted above, the processor 516 is programmed with functionality to display an interface through which a user can input a search request 515 for a particular event. The processor 516 is also connected to a display device 517 which may be a CRT monitor, television, or other display device. The processor 516 would receive the search request, which might include the following terms in a known Boolean structure: "Bob AND Mary AND Kitchen AND stock", in step 5A. These terms would then be matched against the index files 519 stored in the storage device 518 to find the index files that best match the request criteria, in step 5B. Once a match or set of matches is returned to the user, the user can select one of the events identified to be returned to the display, in step 5C. In step 5D, the processor then retrieves the event and plays it on the display.

In an alternate embodiment, the video segments of the data are used to identify persons captured by the recording devices in real-time. With reference to Fig. 6, a flow diagram of a process for controlling and providing or denying access to various home appliances is shown. In this embodiment, the network as shown in Fig. 1 is interconnected to various home appliances, as shown in Figure 1, and the processor is programmed to interact with microprocessors installed in the appliances.

Although the following process is described in connection with the use of a home computer, it is to be understood that one skilled in the art could provide similar functionality for any of the appliances commonly found in the home or office. For the purpose of this example, it is assumed that a recording device (e.g., a video camera) is positioned so as to record the face of the subject trying to the access the appliance. In step

602, the recording device captures a shot of the face of the subject. The shot is then passed to the processing engine in step 604. In step 606, the processing engine uses a face detection technique to analyzed and determine the identity of the individual. To improve the accuracy of the system, a voice recognition technique as earlier described may also be used in combination with the face detection technique. If the individual's face matches one of the faces for which access is to be granted, in step 608, then the processing engine grants access to the computer system, in step 610. If not, then access is denied, in step 612. As such, the individual's face acts as a login or password. Alternatively, where the recording device is a microphone or other audio capture device, a voice recognition system could be used to identify an individual and provide or deny access. Such a system would operate substantially as described above.

With reference back to Fig. 1, according to an embodiment of the present invention, the recording system 10 can constantly record the actions of subjects in the environment 24 hours a day, 7 days a week. In any given day, for example, the recording system 10 may record and identify any number of events or individual actions performed by a particular subject. By identifying the actions, the probabilistic engine can identify those actions which happen repetitively throughout the day or at similar times from day to day. For instance, each night before the subjects go to bed, they may lock the front and back doors of the environment. After several times, the probabilistic engine will identify that this action is performed at night on each day. Thus, the processing system 16 can be programmed to respond to the identified actions in any number of ways, including reminding the subjects to perform the task or actually performing the task for the subjects. By way of non-limiting example, the processing system 16 can be connected to and programmed to operate the electrical systems of the house. Thus, the processing system 16 can turn off the lights when all of the subjects go to bed at night.

In yet another embodiment, the recording device 12, such as a video camera, can be positioned at the front door of the environment 50 to record subjects that approach the door. The recording device 12 can take a snapshot of person(s) visiting the environment and then notify the owner of the environment that a particular person stopped by. This may be done by sending an e-mail to the user at work or storing the snapshot image for later retrieval by the user. The recording device 12 at the front door can also identify a dangerous event when a child member of the environment 50 returns home at an unusual time. For instance, when a child comes home sick from school early, the recording device 12 can record the time and an image of the child returning home so that a parent can be notified of this unusual (and

potential dangerous) event. Again, the snapshot and time stamp can be e-mailed to the parent or communicated in any other way using mobile devices, such as wireless phones or PDAs.

As mentioned earlier, the system can also be used to broadcast content throughout the environment. For instance, a user may wish to listen to an audio book without having to carry a cassette player and headphones with them wherever they travel within the environment. Thus, the sensors or recording devices 12 of the recording system 10 can broadcast the audio book through the speakers interconnected with the system in a particular room in which the subject is located. As the subject moves about the environment, the broadcast audio signal can be sent to those speakers that are in close proximity to the subject. By way of example, if the subject is in the kitchen cooking dinner, the speakers in the kitchen would be active. When the subject moved from the kitchen to the dinning room to eat dinner, the speakers in the dinning room would be activated.

In yet another embodiment, the passive recording system can be used as a monitoring or security system. In such a system, the recording devices are preferably equipped with motion detectors to detect motion and to begin recording upon the appearance of a subject in the field of view of the recording device. If the system is armed and motion is detected, the recording device would record a shot of the subject's face. Then, using a face detection technique, the subject's face could be matched against a database that contains the faces of the individuals that live in the home or work at the office. If a match is not made, then an alarm can be setoff or the proper authorities notified of a possible intrusion. Because the system of the present invention combines both motion detection and face detection, the system is less likely to be falsely setoff by the family dog or other non-intrusive movement.

While the invention has been described in connection with preferred embodiments, it will be understood that modifications thereof within the principles outlined above will be evident to those skilled in the art and thus, the invention is not limited to the preferred embodiments but is intended to encompass such modifications.

CLAIMS:

1. A method of passively recording and indexing events in an operating environment having at least one recording device connected to a network, the network being interconnected to a processor and a storage device, the method comprising:

recording video captured by the recording device;

segmenting the video into at least a video segment and an audio segment;

analyzing the video and audio segments to determine characteristics of the video;

categorizing a portion of the video according to predefined indicia;

associating the characteristics with the analyzed portion of the video; and

storing the video along with the associated category and characteristics on the storage device.

2. The method of claim 1, wherein the segmenting of the video further comprises generating a text transcript of the video.

3. The method of claim 2, further comprising analyzing the text transcript to determine whether a term is used repeatedly.

4. The method of claim 3, wherein the associating further comprises associating the terms used repeatedly with the video.

5. The method of claim 1, wherein a plurality of recording devices are connected to the network.

6. The method of claim 1, wherein the recording device is a video camera.

7. The method of claim 1, wherein the characteristics of the video include a plurality of visual features.

8. The method of claim 1, wherein the analyzing of the video segment further comprises using face detection to identify subjects.

9. The method of claim 1, wherein the processor is connected to a display device and the method further comprises:

receiving a request for a portion of the video;

matching the request to the category and characteristics associated with the video;

displaying the portion of the video matching the request.

10. An adaptive environment system, comprising:

a processing system connectable to a network, the network comprising one or more interconnected sensors, the processing system comprising a computer readable medium comprising computer code for instructing one or more processors to:

receive recorded data from the one or more sensors connectable to the processing system;

analyze the recorded data to identify an event occurring in the recorded data;

determine whether a response to the identified event is appropriate; and

when a response is appropriate generate a signal associated with the response.

11. The system of claim 10, further comprising a storage device communicatively connected to the processing system and further comprising computer code for instructing the one or more processors to:

de-mix the recorded data into at least a video segment and an audio segment;

perform a probabilistic analysis of the video and audio; and

calculate a probability of the recorded data falling within a category.

12. The system of claim 11, wherein the recorded data is archived in the storage device.

13. The system of claim 10, wherein the computer code comprises a probabilistic engine for analyzing the recorded data.

14. The system of claim 13, wherein the probabilistic engine uses a Bayesian approach.

15. The system of claim 10, wherein when the event identified is a dangerous event, the response is to notify a designated person.

16. The system of claim 10, wherein when the event identified is an energy saving event, the response is to control an appliance interconnected to the network.

17. The system of claim 10, wherein when the event identified is a suggestion event, the response is to transmit a message to a user.

18. The system of claim 10, further comprising computer code for instructing the one or more processors to:

create an index of the recorded data;

store the index in an index file; and

store the recorded data along with the index file on a storage device.

19. The system of claim 18, wherein the processing system is created to receive a search request from a user and further comprising computer code for instructing the one or more processors to:

match a parameter of the search request to a portion of the index file; and

return a portion of the recorded data corresponding to the section of the index file matching the parameter of the search request.

20. The system of claim 10, wherein the processing system is programmed to analyze an identity of a recorded subject and perform an action if the recorded subject is unrecognized.

21. The system of claim 20, wherein the action is at least one of:

setting off an alarm,

notifying law enforcement authorities, and

notifying a designated person.

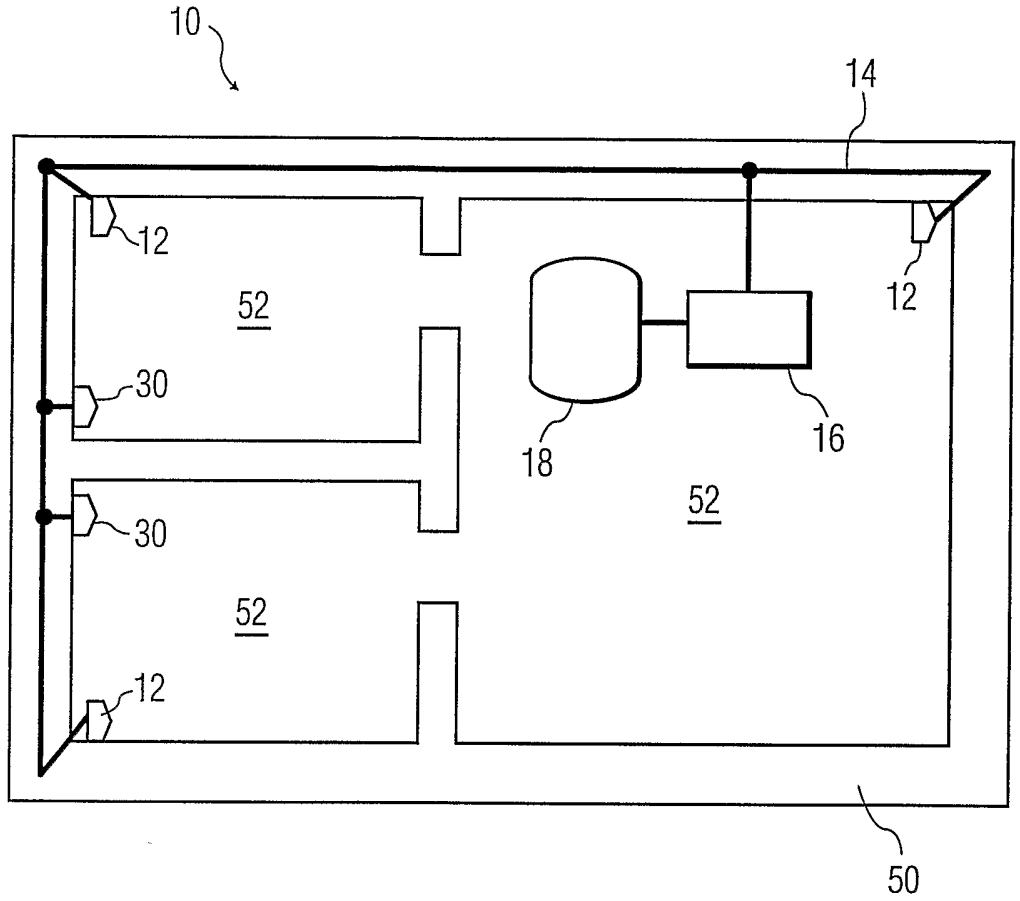


FIG. 1

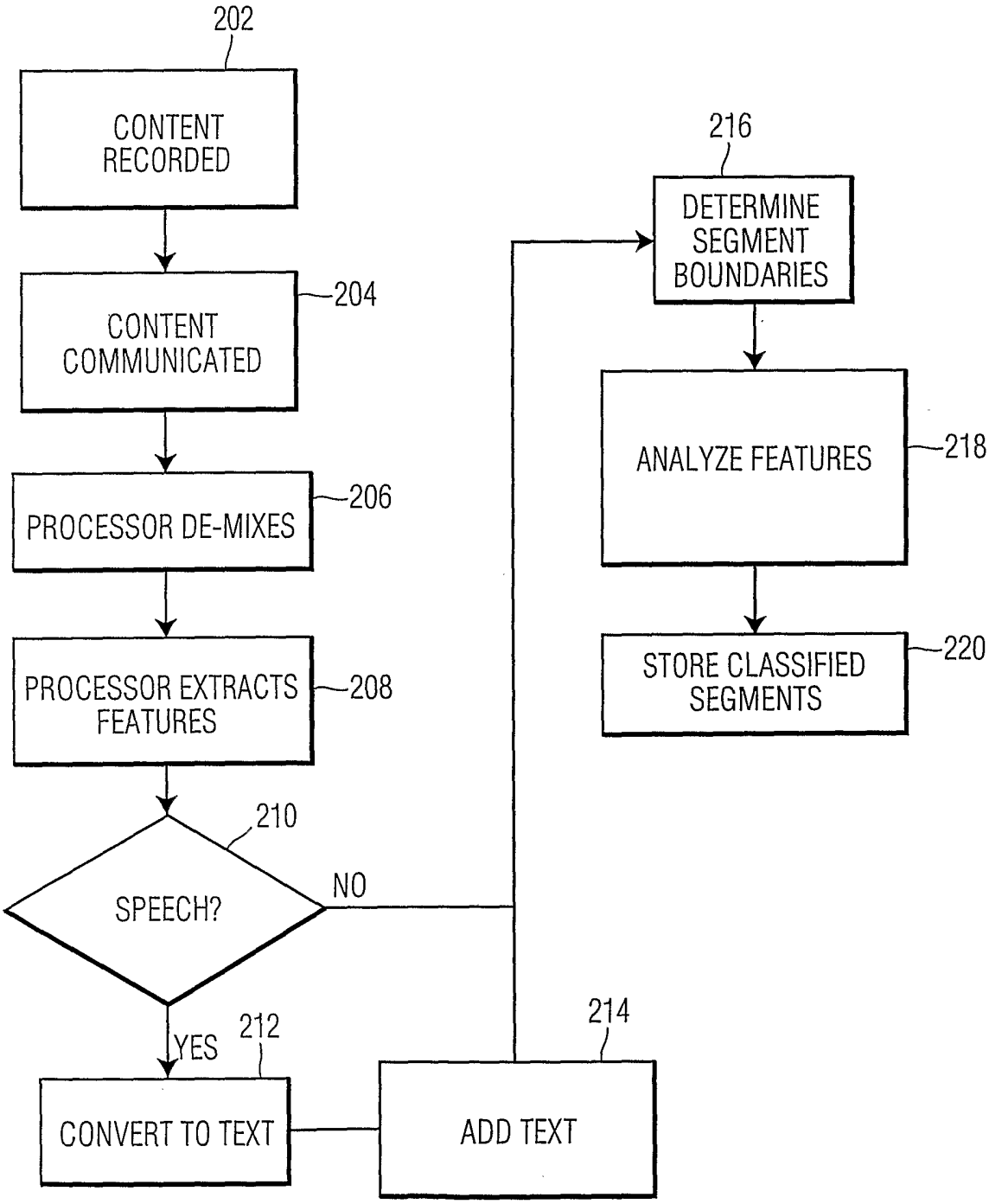

202
CONTENT RECORDED
204
CONTENT COMMUNICATED
206
PROCESSOR DE-MIXES
208
PROCESSOR EXTRACTS FEATURES
210
SPEECH?
NO
YES
212
CONVERT TO TEXT
214
ADD TEXT
216
DETERMINE SEGMENT BOUNDARIES
218
ANALYZE FEATURES
220
STORE CLASSIFIED SEGMENTS

FIG. 2

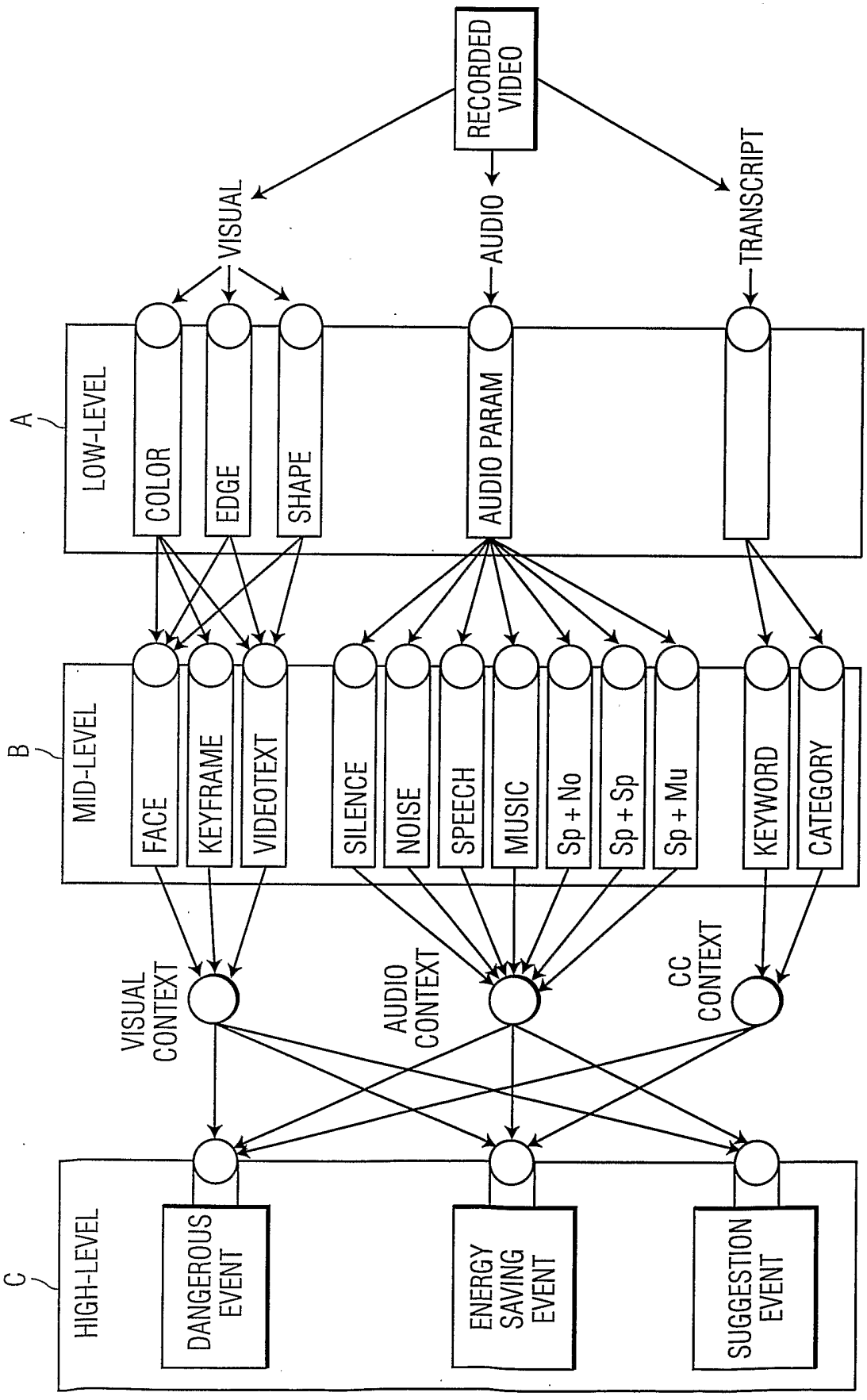

RECORDED VIDEO
VISUAL
AUDIO
TRANSCRIPT
A
LOW-LEVEL
COLOR
EDGE
SHAPE
AUDIO PARAM
B
MID-LEVEL
FACE
KEYFRAME
VIDEOTEXT
SILENCE
NOISE
SPEECH
MUSIC
Sp + No
Sp + Sp
Sp + Mu
KEYWORD
CATEGORY
VISUAL CONTEXT
AUDIO CONTEXT
CC CONTEXT
C
HIGH-LEVEL
DANGEROUS EVENT
ENERGY SAVING EVENT
SUGGESTION EVENT

FIG. 3

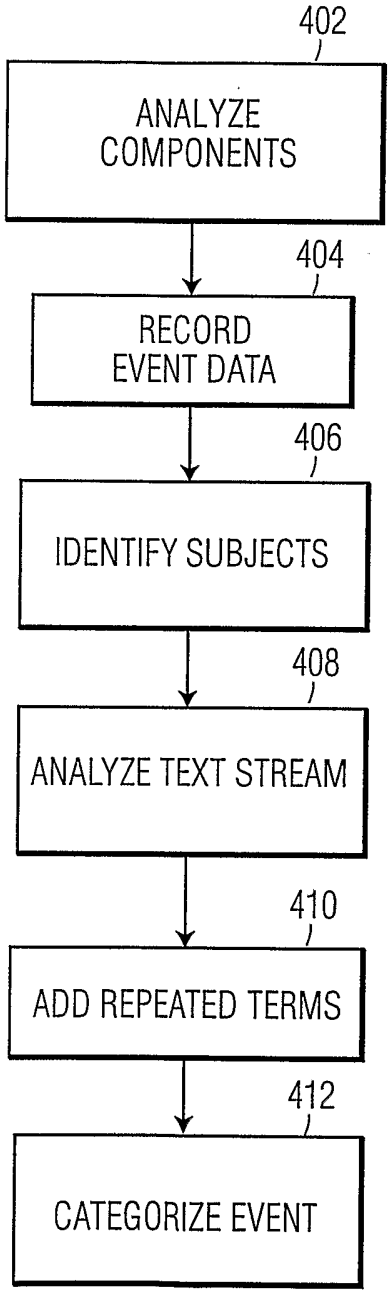

402
ANALYZE COMPONENTS
404
RECORD EVENT DATA
406
IDENTIFY SUBJECTS
408
ANALYZE TEXT STREAM
410
ADD REPEATED TERMS
412
CATEGORIZE EVENT

FIG. 4

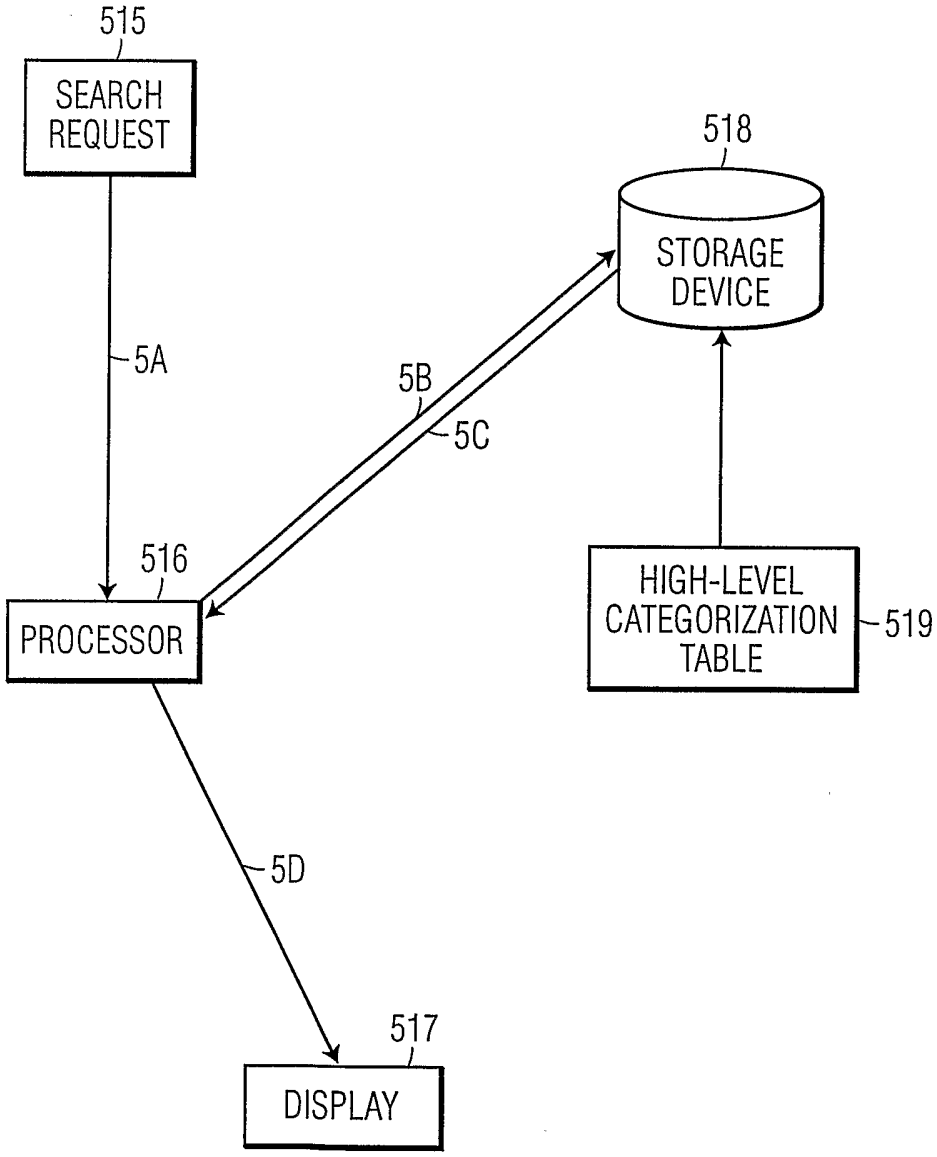

515
SEARCH REQUEST
518
STORAGE DEVICE
5A
5B
5C
516
PROCESSOR
HIGH-LEVEL CATEGORIZATION TABLE
519
5D
517
DISPLAY

FIG. 5

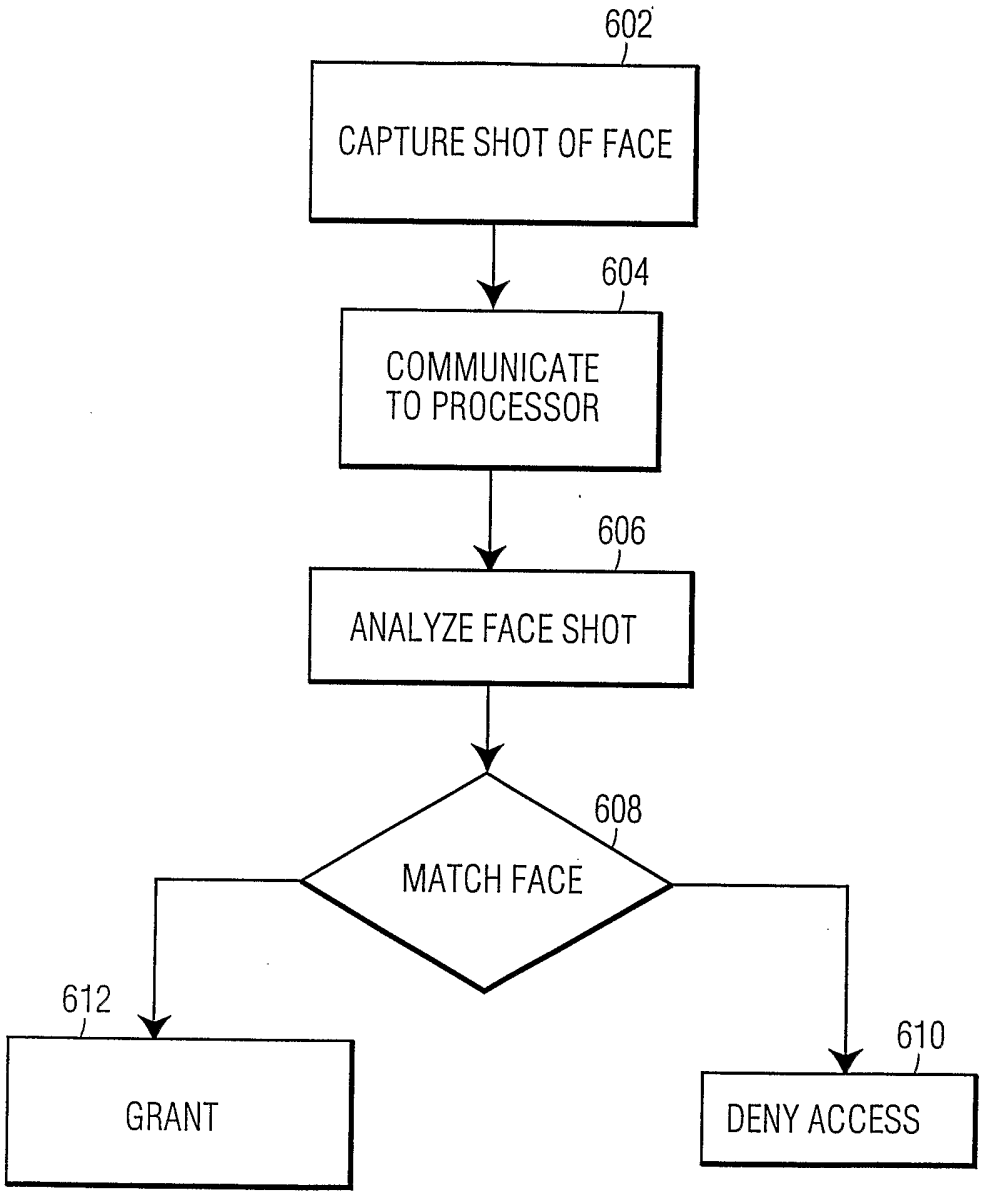

602
CAPTURE SHOT OF FACE
604
COMMUNICATE TO PROCESSOR
606
ANALYZE FACE SHOT
608
MATCH FACE
612
GRANT
610
DENY ACCESS

FIG. 6