

quently than sound samples, a sound-based time stamp generally provides a time reference with higher resolution than does an image-based time stamp. In many cases, the lower resolution of the latter time stamp is of sufficient resolution for purposes of the present invention.

[0064] In one mode of operation, synchronizer 403 issues commands that cause sensors 401 and/or 402 to grab image frames and/or sound samples. Accordingly, the output of synchronizer 403 is frame sync clock 804 and sync clock 504, which are used by frame grabber 503 of sensor 401 and A/D converter 603 of sensor 402, respectively. Synchronizer 403 commands may also cause a time stamp to be attached to each frame or sample. In an alternative embodiment, synchronizer 403 receives notification from sensors 401 and/or 402 that an image frame or a sound sample has been acquired, and attaches a time stamp to each.

[0065] In an alternative embodiment, synchronizer 403 is implemented in software. For example, frame grabber 503 may generate an interrupt whenever it captures a new image. This interrupt then causes a software routine to examine the computer's internal clock, and the time the latter returns is used as the time stamp for that frame. A similar procedure can be used for sound samples. In one embodiment, since the sound samples are usually acquired at a much higher rate than are image frames, the interrupt may be called only once every several sound samples. In one embodiment, synchronizer 403 allows for a certain degree of tolerance in determining whether events in two domains are synchronous. Thus, if the time stamps indicate that the events are within a predefined tolerance time period of one another, they are deemed to be synchronous. In one embodiment, the tolerance time period is 33 ms, which corresponds to a single frame period in a standard video camera.

[0066] In an alternative software implementation, the software generates signals that instruct optical sensor 401 and acoustic sensor 402 to capture frames and samples. In this case, the software routine that generates these signals can also consult the system clock, or alternatively it can stamp sound samples with the number of the image frame being grabbed in order to enforce synchronization. In one embodiment, optical sensor divider 801 and acoustic sensor divider 802 are either hardware circuitry or software routines. Dividers 801 and 802 count pulses from master clock 803, and output a synchronization pulse after every sequence of predetermined length of master-clock pulses. For instance, master clock 803 could output pulses at a rate of 1 MHz. If optical sensor divider 801 controls a standard frame grabber 503 that captures images at 30 frames per second, divider 801 would output one frame sync clock pulse 804 every $1,000,000/30 \approx 33,333$ master-clock pulses. If acoustic sensor 402 captures, say, 8,000 samples per second, acoustic sensor divider 802 would output one sync clock pulse 504 every $1,000,000/8,000 = 125$ master clock pulses.

[0067] One skilled in the art will recognize that the above implementations are merely exemplary, and that synchronizer 403 may be implemented using any technique for providing information relating acquisition time of visual data with that of sound data.

[0068] Referring now to FIG. 9, there is shown an example of an implementation of processor 404 according to one embodiment. Processor 404 may be implemented in software or in hardware, or in some combination thereof.

Processor 404 may be implemented using components that are separate from other portions of the system, or it may share some or all components with other portions of the system. The various components and modules shown in FIG. 9 may be implemented, for example, as software routines, objects, modules, or the like.

[0069] Processor 404 receives sound information 604 and visual information 505, each including time stamp information provided by synchronizer 403. In one embodiment, portions of memory 105 are used as first-in first-out (FIFO) memory buffers 105A and 105B for audio and video data, respectively. As will be described below, processor 404 determines whether sound information 604 and visual information 505 concur in detecting occurrence of an intended user action of a predefined type that involves both visual and acoustic features.

[0070] In one embodiment, processor 404 determines concurrence by determining the simultaneity of the events recorded by the visual and acoustic channels, and the identity of the events. To determine simultaneity, processor 404 assigns a reference time stamp to each of the two information streams. The reference time stamp identifies a salient time in each stream; salient times are compared to the sampling times to determine simultaneity, as described in more detail below. Processor 404 determines the identity of acoustic and visual events, and the recognition of the underlying event, by analyzing features from both the visual and the acoustic source. The following paragraphs describe these operations in more detail.

[0071] Reference Time Stamps: User actions occur over extended periods of time. For instance, in typing, a finger approaches the typing surface at velocities that may approach 40 cm per second. The descent may take, for example, 100 milliseconds, which corresponds to 3 or 4 frames at 30 frames per second. Finger contact generates a sound towards the end of this image sequence. After landfall, sound propagates and reverberates in the typing surface for a time interval that may be on the order of 100 milliseconds. Reference time stamps identify an image frame and a sound sample that are likely to correspond to finger landfall, an event that can be reliably placed in time within each stream of information independently. For example, the vision reference time stamp can be computed by identifying the first image in which the finger reaches its lowest position. The sound reference time stamp can be assigned to the sound sample with the highest amplitude.

[0072] Simultaneity: Given two reference time stamps from vision and sound, simultaneity occurs if the two stamps differ by less than the greater of the sampling periods of the vision and sound information streams. For example, suppose that images are captured at 30 frames per second, and sounds at 8,000 samples per second, and let t_v and t_s be the reference time stamps from vision and sound, respectively. Then the sampling periods are 33 milliseconds for vision and 125 microseconds for sound, and the two reference time stamps are simultaneous if $|t_v - t_s| \leq 33$ ms.

[0073] Identity and Classification: Acoustic feature computation module 901 computes a vector a of acoustic features from a set of sound samples. Visual feature computation module 902 computes a vector v of visual features from a set of video samples. Action list 905, which may be stored in memory 105C as a portion of memory 105, describes a set