

[0108] Similar histograms v^1, K, v^M are pre-computed for M (ranging, in one embodiment, between 2 and 10) hand configurations of interest, corresponding to at most M different commands.

[0109] At operation time, reference time stamps are issued whenever the value of

$$\min_m \|v - v^m\|$$

[0110] falls below a predetermined threshold, and reaches a minimum value over time. The value of m that achieves this minimum is the candidate gesture for the vision system.

[0111] Suppose now that at least some of the stored vectors v^m correspond to gestures emitting a sound, such as a snap of the fingers or a clap of hands. Then, acoustic feature computation module 901 determines the occurrence of, and reference time stamp for, a snap or clap event, according to the techniques described above.

[0112] Even if the acoustic feature computation module 901 or the visual feature computation module 902, working in isolation, would occasionally produce erroneous detection results, the present invention reduces such errors by checking whether both modules agree as to the time and nature of an event that involves both vision and sound. This is another instance of the improved recognition and interpretation that is achieved in the present invention by combining visual and auditory stimuli. In situations where detection in one or the other domain by itself is insufficient to reliably recognize a gesture, the combination of detection in two domains can markedly improve the rejection of unintended gestures.

[0113] The techniques of the present invention can also be used to interpret a user's gestures and commands that occur in concert with a word or brief phrase. For example, a user may make a pointing gesture with a finger or arm to indicate a desired direction or object, and may accompany the gesture with the utterance of a word like "here" or "there." The phrase "come here" may be accompanied by a gesture that waves a hand towards one's body. The command "halt" can be accompanied by an open hand raised vertically, and "good bye" can be emphasized with a wave of the hand or a military salute.

[0114] For such commands that are simultaneously verbal and gestural, the present invention is able to improve upon conventional speech recognition techniques. Such techniques, although successful in limited applications, suffer from poor reliability in the presence of background noise, and are often confused by variations in speech patterns from one speaker to another (or even by the same speaker at different times). Similarly, as discussed above, the visual recognition of pointing gestures or other commands is often unreliable because intentional commands are hard to distinguish from unintentional motions, or movements made for different purposes.

[0115] Accordingly, the combination of stimulus detection in two domains, such as sound and vision, as set forth herein, provides improved reliability in interpreting user gestures when they are accompanied by words or phrases. Detected stimuli in the two domains are temporally matched in order to classify an input event as intentional, according to techniques described above.

[0116] Recognition function 903 $r_c(a, v)$ can use conventional methods for speech recognition as are known in the art, in order to interpret the acoustic input a, and can use conventional methods for gesture recognition, in order to interpret visual input v. In one embodiment, the invention determines a first probability value $p_a(u)$ that user command u has been issued, based on acoustic information a, and determines a second probability value $p_v(u)$ that user command u has been issued, based on visual information v. The two sources of information, measured as probabilities, are combined, for example by computing the overall probability that user command u has been issued:

$$p = 1 - (1 - p_a(u))(1 - p_v(u))$$

[0117] p is an estimate of the probability that both vision and hearing agree that the user intentionally issued gesture u. It will be recognized that if $p_a(u)$ and $p_v(u)$ are probabilities, and therefore numbers between 0 and 1, then p is a probability as well, and is a monotonically increasing function of both $p_a(u)$ and $p_v(u)$. Thus, the interpretation of p as an estimate of a probability is mathematically consistent.

[0118] For example, in the example discussed with reference to FIG. 12, the visual probability $p_v(u)$ can be set to

$$p_v(u) = K_v e^{-(v - v^m)^2}$$

[0119] where K_v is a normalization constant. The acoustic probability can be set to

$$p_a(u) = K_a e^{-\alpha^2}$$

[0120] where K_a is a normalization constant, and α is the amplitude of the sound recorded at the time of the acoustic reference time stamp.

[0121] In the above description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the invention. It will be apparent, however, to one skilled in the art that the invention can be practiced without these specific details. In other instances, structures and devices are shown in block diagram form in order to avoid obscuring the invention.

[0122] Reference in the specification to "one embodiment" or "an embodiment" means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the invention. The appearances of the phrase "in one embodiment" in various places in the specification are not necessarily all referring to the same embodiment.

[0123] Some portions of the detailed description are presented in terms of algorithms and symbolic representations of operations on data bits within a computer memory. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated. It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like.