

be a part of a longer gesture required for a corresponding command. To use the sand example, the command for picking up sand may be mapped to a sequence of gestures, such as a hand press touch gesture, followed by a non-touch grasp gesture (or making a fist). The system may detect the hand press touch gesture, and pass through steps 606 and 608, returning to 603 to detect the ensuing grasp gesture before executing the command for picking up the sand.

[0047] In the discussion of FIG. 6 above, the touch and non-touch gesture detections were considered together in step 605, resulting in a determination of a recognized touch and/or non-touch gesture. This “fusion” of touch and non-touch gesture detection can be achieved in a variety of ways, to allow both touch and non-touch gestures to be integrated into a user interface scheme for the system 100. FIG. 7 illustrates one example process for this fusion of step 605.

[0048] In step 701, the process may first align in time the video frames captured by cameras 203 and 204, so that the system compares the same scene from the two cameras. For example, the system may choose video frames taken by the cameras 203, 204 at time 12:01:00. Then, in step 702, the system may use the image from camera 203 to determine whether, from camera 203’s point of view, a predetermined gesture has been detected. This may follow the detection process discussed above in FIG. 5. In doing so, the system may determine not only whether a predetermined gesture’s template matches the image in the frame, but it may also determine a similarity value S_T for the match, and may output (or store) the identification and the similarity value.

[0049] In step 703, the same gesture detection process may be performed using the image data from camera 204, resulting in a gesture identification and similarity value S_H based on the camera 204 image.

[0050] In step 704, the touch and non-touch similarities (S_T , S_H) may be normalized to make comparison easier (e.g., by normalizing them to a standard scale, such as a percentage of a perfect match), and they may then be compared with gesture validation thresholds (T_T , T_H) that may be predetermined for the different cameras 203, 204. The thresholds may determine the minimum level of similarity that will be required for the system to accept the detected gesture as actually having occurred. The actual threshold values may be established through a calibration/training process. For example, the system may initially ask the user to perform one or more predetermined, known gestures (e.g., prompting the user to “please place both hands palm-side down on the display”) to obtain a baseline measurement of a gesture, and then the thresholds may be set a given percentage (e.g., 10%) off, to allow a predetermined deviation in distance, size, etc. This calibration/training may also occur over a period of use, or continually. For example, the system may allow the user to provide feedback indicating whether a particular gesture was accurately detected, and the system may adjust its threshold data to increase (or decrease) the threshold value to accommodate the gesture.

[0051] For example, the system may require a 50% certainty for gestures detected from camera 203, and a 75% certainty for gestures detected from camera 204. If, in step 704, it is determined that neither of the similarities (S_T , S_H) exceeds its corresponding threshold (T_T , T_H) (e.g., $S_T < T_T$ and $S_H < T_H$), then the process may proceed to step 705, and indicate that no suitable touch/non-touch gesture was detected.

[0052] If, however, at least one of the thresholds was met, then the process may proceed to step 706, and determine whether only one of the thresholds was met. If only one was met (e.g., only $S_T > T_T$; or only $S_H > T_H$), then the process may proceed to step 707, in which the gesture identified from the

camera whose threshold was met is output. For example, if only the similarity from camera 203 (S_T) exceeded its threshold ($S_T > T_T$), then the system may generate an output indicating that the gesture identified in step 702 has been detected. Conversely, if the similarity from camera 204 (S_H) was the only one to exceed its threshold (T_H), then the system may generate an output indicating that the gesture identified in step 703 has been detected.

[0053] If, in step 706, it is determined that both thresholds were met, then the process may proceed to step 708, to determine which camera should be believed. If the similarity value from one camera is much stronger than the similarity value from the other camera, then the gesture identification from the first camera is output. This may be implemented by calculating a difference between the similarities (e.g., $|S_T - S_H|$), and setting a differential threshold (T_D) to determine how much stronger one camera’s similarity value must be. For example, if the difference in similarity exceeds the differential threshold ($|S_T - S_H| > T_D$), then in step 709 the gesture identification from the camera having the higher similarity value is output.

[0054] However, in step 708, if the similarity values from the two cameras 203, 204 are close to one another (e.g., difference less than SD), the system may proceed to step 710, and employ a gesture state machine algorithm for determining which gesture identification should control. The gesture state machine algorithm may use the context of the application to determine which gesture detection is to be used. For example, the context information may identify the previous detected gesture, and the determination may compare the identified gestures in step 710 with the previous detected gesture.

[0055] The previous gesture may have associated with it a predetermined prioritized list identifying the likelihood of a subsequent gesture. For example, a template for a sand grasping gesture may indicate that this gesture is more likely to be followed by a sand releasing gesture, and that it is less likely to be followed by a pressing gesture. The system can, in step 710, consult this context information and select the more likely gesture. Other contextual information may be used as well, such as the hand position (from position A, positions B and C are more likely than position D), gesture frequency (in event of a tie in step 710, choose the more common gesture of the two identified gestures), command set (an application may have a subset of commands that are more likely to be used), etc.

[0056] After the detected gesture is output, the process may then terminate (or return to step 606, if the FIG. 7 process is used to implement step 605).

[0057] Although examples of carrying out the features described herein have been described, there are numerous other variations, combinations and permutations of the above described devices and techniques may exist as desired. For example, process steps may be rearranged, combined, omitted, interrupted, etc.; variable values may be changed, etc. The various structures and systems described herein may also be subdivided, combined, or varied as desired. For example, the touch-based system and non-touch based system need not be wholly separate systems, and may instead share components, such as cameras, display screens, processor capacity, memory, computer code, etc. Components and process steps may also be omitted. For example, the display screen may, if desired, be replaced with a simple surface, such as a touch pad.

[0058] The above description and drawings are illustrative only. The features herein are not limited to the illustrated embodiments, and all embodiments of the invention need not necessarily achieve all of the advantages or purposes, or possess all characteristics, identified herein.