



- (51) **International Patent Classification:**
G06K 9/34 (2006.01) *G06T 1/00* (2006.01)
- (21) **International Application Number:**
PCT/CN2014/085381
- (22) **International Filing Date:**
28 August 2014 (28.08.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant:** QUALCOMM INCORPORATED [US/US];
5775 Morehouse Drive, San Diego, California 92121-1714 (US).
- (72) **Inventors; and**
- (71) **Applicants (for US only):** ZHONG, Xin [US/US]; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). GAO, Dashan [CN/US]; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). SUN, Yu [CN/CN]; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). QI, Yingyong [US/US]; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). ZHENG, Baozhong [CN/US]; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). BOSCH RUIZ, Marc [ES/US]; 5775 Morehouse Drive, San Diego, California 92121-1714 (US). KAMATH, Nagendra [IN/US]; 5775 Morehouse Drive, San Diego, California 92121-1714 (US).
- (74) **Agent:** LEE AND LI - LEAVEN IPR AGENCY LTD.;
Unit 2202, Tower A, Beijing Marriott Center, No.7 Jian

Guo Men South Avenue, Dongcheng District, Beijing 100005 (CN).

- (81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

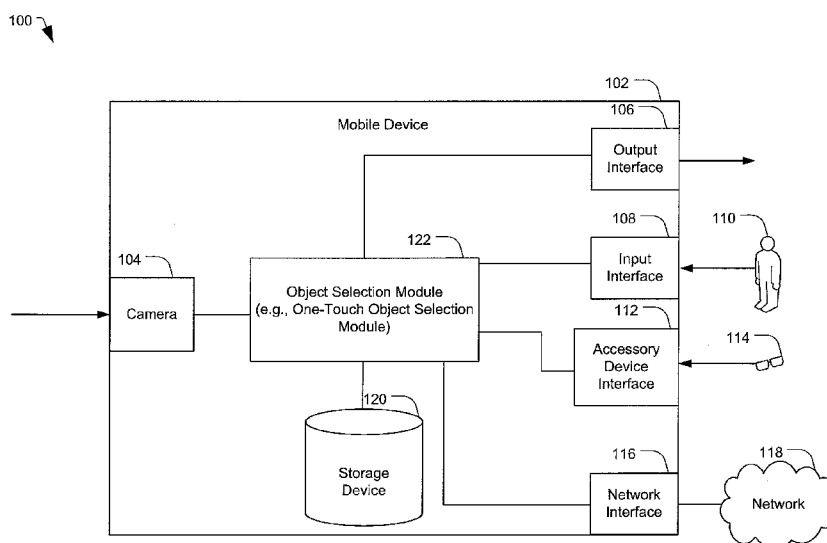
Declarations under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

— with international search report (Art. 21(3))

(54) **Title:** OBJECT SELECTION BASED ON REGION OF INTEREST FUSION

**FIG. 1**

(57) **Abstract:** A method includes receiving a user input (e. g., a one-touch user input), performing segmentation to generate multiple candidate regions of interest (ROIs) in response to the user input, and performing ROI fusion to generate a final ROI (e. g., for a computer vision application). In some cases, the segmentation may include motion-based segmentation, color-based segmentation, or a combination thereof. Further, in some cases, the ROI fusion may include intraframe (or spatial) ROI fusion, temporal ROI fusion, or a combination thereof.

OBJECT SELECTION BASED ON REGION OF INTEREST FUSION

I. Field

[0001] The present disclosure is generally related to object selection.

II. Description of Related Art

[0002] Advances in technology have resulted in smaller and more powerful computing devices. For example, there currently exist a variety of portable personal computing devices, including wireless telephones such as mobile and smart phones, tablets and laptop computers that are small, lightweight, and easily carried by users. These devices can communicate voice and data packets over wireless networks. Further, many such devices incorporate additional functionality such as a digital still camera, a digital video camera, a digital recorder, and an audio file player. Also, such devices can process executable instructions, including software applications, such as a web browser application, that can be used to access the Internet. As such, these devices can include significant computing capabilities.

[0003] Object selection is associated with various computer vision (CV) use cases, and various types of inputs may be associated with initial object selection. To illustrate, for example CV use cases such as object tracking, object recognition, or augmented reality, a user may perform an initial selection of the object. However, object selection may be cumbersome in some cases, as the object may be defined by a two-dimensional mask or by a bounding box. As an illustrative example, on a touch screen of mobile communication device, this selection may be performed by “drawing” the bounding box on the touch screen. That is, the user may use a two-finger or a one-finger “draw” to cross the object by drawing a line that defines the bounding box, which may be unintuitive and imprecise. This method of object selection process may be more difficult for small objects or moving objects.

III. Summary

[0004] The present disclosure describes an object selection scheme that may be used in various applications, including object tracking, object recognition, reality augmentation,

and scene analysis, among other alternatives. In some cases, it may be difficult for a user to define a bounding box using a multi-touch input (e.g., a one-finger draw or a two-finger draw). For example, it may be difficult to define a bounding box around a moving object or around a small object using a multi-touch input. Accordingly, the present disclosure describes a method of object selection that may provide an improved user experience by generating an object bounding box responsive to a user input other than a multi-touch input (e.g., a one-touch user input).

[0005] In a particular example, a method of object selection is disclosed. The method includes receiving a user input, performing segmentation to generate multiple candidate regions of interest (ROIs), and performing ROI fusion to generate a final ROI (e.g., for a computer vision application). In some cases, the segmentation may include motion-based segmentation, color-based segmentation, or a combination thereof. Further, in some cases, the ROI fusion may include intraframe (or spatial) ROI fusion, temporal ROI fusion, or a combination thereof.

[0006] In another particular example, an apparatus for object selection is disclosed. The apparatus includes a processor, an input device to receive a user input, and a video encoder to generate a motion vector field responsive to the user input. The apparatus further includes a segmentation component executable by the processor to perform segmentation to generate multiple candidate regions of interest. The apparatus also includes a fusion component executable by the processor to perform region of interest fusion to generate a final region of interest.

[0007] In another particular example, an apparatus for object selection is disclosed. The apparatus includes means for receiving a user input and means for generating a motion vector field responsive to the user input. The apparatus further includes means for performing segmentation to generate multiple candidate regions of interest. The apparatus also includes means for performing region of interest fusion to generate a final region of interest.

[0008] In another particular example, a computer-readable storage device is disclosed. The computer-readable storage device stores instructions that are executable by a processor to perform various operations associated with a method of object selection. The operations may include receiving a user input, performing segmentation to generate

multiple candidate regions of interest (ROIs), and performing ROI fusion to generate a final ROI.

[0009] One particular advantage provided by at least one of the disclosed examples is an improved user experience with respect to object selection, particularly in the context of selection of small objects or moving objects.

[0010] Other aspects, advantages, and features of the present disclosure will become apparent after review of the entire application, including the following sections: Brief Description of the Drawings, Detailed Description, and the Claims.

IV. Brief Description of the Drawings

[0011] FIG. 1 is a block diagram that illustrates a particular embodiment of a system that is operable to perform one-touch object selection;

[0012] FIG. 2 is a diagram of a particular illustrative embodiment of a method of one-touch object selection;

[0013] FIG. 3 is a block diagram of a particular illustrative embodiment of example computer vision (CV) applications that may utilize the results of an initial object selection;

[0014] FIG. 4 is a block diagram of a particular illustrative embodiment of a method of one-touch object selection that includes motion-based segmentation to generate candidate regions of interest (ROIs) and intraframe (or spatial) ROI fusion to generate a final ROI (e.g., for an object tracking application);

[0015] FIG. 5 is a block diagram of a particular illustrative embodiment of a method of error handling responsive to the final ROI of FIG. 4 not satisfying one or more ROI criteria;

[0016] FIG. 6 is a block diagram of a particular illustrative embodiment of a method of one-touch object selection that includes color-based segmentation;

[0017] FIG. 7 is a flowchart of a particular illustrative embodiment of a method of one-touch object selection that includes color-based segmentation to generate candidate

ROIs and temporal ROI fusion to generate a final ROI (e.g., for an object tracking application);

[0018] FIG. 8 is a flowchart of a particular illustrative embodiment of a method that may include both motion-based segmentation and color-based segmentation to generate a final ROI; and

[0019] FIG. 9 is a block diagram that illustrates a particular embodiment of a wireless device configured to perform one-touch object selection.

V. Detailed Description

[0020] Referring to FIG. 1, a particular illustrative embodiment of a system that is operable to perform one-touch object selection is disclosed and generally designated 100. The system 100 includes a mobile device 102. The mobile device 102 may be a mobile phone, a music player, a video player, an entertainment unit, a navigation device, a communications device, a personal digital assistant (PDA), a computer, or any other mobile computing device. The mobile device 102 includes a camera 104. The camera 104 may be configured to capture and output still images and videos. The mobile device 102 includes an output interface 106. The output interface 106 may be configured to communicate with a display device, such as a liquid crystal display (LCD), a light emitting diode (LED) display, or any other display device. In a particular embodiment, the output interface 106 outputs a graphical user interface (GUI). The mobile device 102 further includes an input interface 108. The input interface 108 may include a touch screen, any other type of input device, or any combination thereof. In particular embodiments, the input interface 108 may be configured to receive input from a user 110 (e.g., input responsive to a GUI output by the output interface 106).

[0021] The mobile device 102 may further include an accessory device interface 112. In a particular embodiment, the accessory device interface 112 receives input from an accessory device 114. In a particular embodiment, the accessory device 114 includes a camera. The input received from the accessory device 114 may include image or video data. In a particular embodiment, the accessory device 114 may be embedded in a user wearable accessory, such as eyeglasses or jewelry.

[0022] The mobile device 102 may further include a network interface 116 configured to communicate with a network 118. The network interface 116 may include an Ethernet interface, an 802.11 (WiFi) interface, a Long Term Evolution (LTE) interface, a Code Division Multiple Access (CDMA) interface, a Time Division Multiple Access (TDMA) interface, an 802.16 (WiMAX) interface, any other wired or wireless network interface, or any combination thereof.

[0023] The mobile device 102 further includes a storage device 120. The storage device 120 may include a solid state drive, a hard disk drive, an optical drive, or any other type of computer readable storage medium or device. The storage device 120 may store images and videos (e.g., images and videos that are captured by the camera 104, downloaded by the mobile device 102 via the network interface 116, etc.).

[0024] An object selection module 122 (e.g., a one-touch object selection module) may be implemented in software (e.g., instructions stored in a memory of the mobile device 102 that are executable by a processor of the mobile device 102). Alternatively, all or part of the object selection module 122 may be implemented in hardware. The object selection module 122 may receive, via user input, selections of one or more objects included (e.g., depicted) in an image or a frame of video. In some embodiments, the object selection module 122 may be configured to perform object selection in response to a one-touch input received from the user 110. Examples of operation of the system 100 are further described with reference to FIGS. 2-8.

[0025] Referring to FIG. 2, a particular illustrative embodiment of a method of one-touch object selection is disclosed and generally designated 200. FIG. 2 illustrates that a user may select an object 202 via a one-touch input 204, and an object bounding box 206 may be identified responsive to the one-touch input 204.

[0026] One-touch object selection may be useful in various computer vision (CV) applications. As an illustrative, non-limiting example, a multi-touch input to define a bounding box may be cumbersome or imprecise in an object tracking application. In order to define a bounding box using a multi-touch input, the user may cross an object by drawing a line using a one finger draw or a two finger or a two finger draw. Such a bounding box may be imprecise. For example, the user may select more or less of the image for tracking than desired. Further, in some cases it may be difficult for the user

to define a bounding box around a moving object (e.g., a fast moving car) or around a small object (e.g., a particular soccer player on a soccer field). Accordingly, generating the object bounding box 206 to select the object 202 in response to the one-touch input 204 may provide an improved user experience.

[0027] Referring to FIG. 3, multiple example computer vision (CV) use cases associated with object selection (e.g., in response to a one-touch input) are illustrated and generally designated 300.

[0028] FIG. 3 illustrates that an initial object selection 302 may be associated with various applications. For example, the initial object selection 302 may be based on input 301. The input 301 may include user input, such as a one-touch input on a touch screen (e.g., the one-touch input 204 illustrated in FIG. 2). However, it will be appreciated that there may be multiple ways for a user to make an initial selection of an object. Examples of alternative user inputs may include one or more gestures, one or more eye movements, or ultrasound sensor input based on detection of a stylus or other device in the possession of the user. Alternatively, various CV-based automatic object detection mechanisms may be employed for initial object selection.

[0029] FIG. 3 further illustrates that the initial object selection 302 may be useful in various applications, including an object tracking application 304, an object recognition application 306, a reality augmentation application 308, or a scene analysis application 310, among other alternatives. In the example image associated with the object tracking application 304, the object being tracked includes a moving car 312. In the example image associated with the object recognition application 306, four objects are identified, including a human 314, a plane 316, a car 318, and an animal 320. In the example image associated with the reality augmentation application 308, information 322 associated with a particular location is provided (e.g., an address of a building or an indication that a monument is located near the building). In the example image associated with the scene analysis application 310, individual soccer players on a soccer field may be identified by a different bounding region 324.

[0030] Referring to FIG. 4, a particular illustrative embodiment of a method of motion-based segmentation for object selection is disclosed and generally designated 400. In the example illustrated in FIG. 4, object selection using motion-based segmentation may

be responsive to a one-touch input 402 (e.g., responsive to a single user touch of a person in an image). In FIG. 4, the one-touch input 402 is represented as a white dot on the back of the running child.

[0031] Responsive to the one-touch input 402, motion may be detected based on at least two video frames. FIG. 4 illustrates an example in which a sequence of video frames 404 including a first video frame 406, a second video frame 408, and a third video frame 410 are used for motion field generation 412. However, it will be appreciated that an alternative number of video frames may be used for motion field generation 412. In some cases, a video encoder (e.g., video encoding hardware) may be used for global/local motion estimation 414. In some cases, the video encoder may estimate motion using a subset of video encoding stages associated with motion estimation without performing other video encoding stages that are not associated with motion estimation.

[0032] FIG. 4 illustrates an example of a motion vector field 416 generated by the video encoder. In some cases, the motion vector field 416 may represent a dense motion vector field (e.g., a motion vector for every 8x8 block of pixels in a frame). While the motion vector field 416 is illustrated in a grayscale format in FIG. 4, the motion vector field 416 may include one or more colors. While the motion vector field 416 may be noisy, the motion vector field 416 of FIG. 4 illustrates that a moving person is discernible. For the global/local motion estimation 414, further processing of the motion vector field 416 may be performed. For example, FIG. 4 illustrates a first grayscale image 418 that represents X direction (horizontal) motion in the motion vector field 416 and a second grayscale image 420 that represents Y direction (vertical) motion in the motion vector field 416. In the particular example illustrated in FIG. 4, the first grayscale image 418 represents the results of applying an X direction median filter to the motion vector field 416, while the second grayscale image 420 represents the results of applying a Y direction median filter to the motion vector field 416. In alternative embodiments, one or more different filters or sets of filters may be employed to further process the motion vector field 416.

[0033] In a particular embodiment, global motion estimation may include determining a median of all motion in both the X direction and the Y direction. Alternatively, other

methods of global motion estimation may be employed. For example, an image may be divided into multiple regions (e.g., 8x8 pixel squares), a median of motion may be obtained for each region, and global motion may be estimated based on a median of the individual medians from the multiple regions. In a particular embodiment, local motion estimation may include determining local motion vectors in individual portions of the image (e.g., in individual 8x8 pixel squares).

[0034] In the example illustrated in FIG. 4, the one-touch input 402 may be used to separate local motion from global motion. That is, the one-touch input 402 may be associated with X and Y coordinates in the motion vector field 416, and these X and Y coordinates may represent a starting location to be used as a first seed 422 for region growing. In FIG. 4, the first seed 422 is represented by a dot, and a first region growing operation performed based on the first seed 422 results in a first region of interest (ROI) 424 (also referred to herein as a bounding box).

[0035] In some cases, a bounding box that is generated by region growing based on the one-touch input 402 may not satisfy a bounding box size threshold associated with an object tracking application (e.g., the object tracking application 304 of FIG. 3). As another example, a user may not accurately select a particular object via the one-touch input 402. For example, it may be difficult for the user to select small objects (e.g., the soccer player 324 in FIG. 2) and/or fast moving objects (e.g., the moving car 312 in FIG. 2). Accordingly, while the one-touch input 402 may provide a starting point for region growing, FIG. 4 illustrates a particular embodiment of segmentation by region growing 426 that uses one or more alternative seeds for region growing. It will be appreciated that segmentation may be performed using various methods (e.g., thresholding, grabcut, etc.) that may not use one or more seed points for segmentation.

[0036] FIG. 4 illustrates that multiple candidate regions of interest (ROIs) 428 may be generated by region growing from multiple seeds (or multiple sets of seeds). A first candidate ROI includes the first ROI 424 that is generated by region growing using the one-touch input 402 as the first seed 422. FIG. 4 further illustrates a particular example in which four other seeds are used for region growing. However, it will be appreciated that an alternative number of seeds (or sets of seeds) may be used for the segmentation by region growing 426, resulting in an alternative number of candidate ROIs. In the

example of FIG. 4, the four other seeds are neighboring X,Y coordinates with respect to the X,Y coordinates of the first seed 422. In some cases, neighboring X,Y coordinates may include coordinates that are offset by n pixels (in a positive or negative direction), where n may be an integer that is fixed (e.g., 1) or programmable. As an illustrative, non-limiting example, region growing based on a second seed with alternative X,Y coordinates (e.g., X-1, Y+1) may result in a second candidate ROI 430. As further examples, region growing based on a third seed with alternative coordinates (e.g., X+1, Y+1) may result in a third candidate ROI 432, region growing based on a fourth seed with alternative coordinates (e.g., X-1, Y-1) may result in a fourth candidate ROI 434, and region growing based on a fifth seed with alternative coordinates (e.g., X+1, Y-1) may result in a fifth candidate ROI 436.

[0037] FIG. 4 further illustrates that intraframe ROI fusion 438 (also referred to herein as spatial ROI fusion) may be performed on at least a subset of the candidate ROIs 428 in order to generate a final ROI 440. That is, the individual candidate ROIs 424, 430, 432, 434, and 436 represent ROIs that are generated by individual region growing operations performed based on different seeds, and the final ROI 440 represents a fused result of the individual region growing operations. In the particular example illustrated in FIG. 4, the final ROI 440 is defined by a maximum X span and a maximum Y span of the individual candidate ROIs 424, 430, 432, 434, and 436. Alternatively, one or more of the candidate ROIs 424, 430, 432, 434, and 436 may be discarded, and intraframe ROI fusion 438 may be performed on a subset of the candidate ROIs 424, 430, 432, 434, and 436. To illustrate, one or more of the five candidate ROIs 424, 430, 432, 434, and 436 may be discarded when they do not satisfy a size threshold (e.g., the ROI may be too small for object tracking). As another example, one or more of the five candidate ROIs 424, 430, 432, 434, and 436 may be discarded when they exceed a size threshold (e.g., the ROI may be too large for object tracking). That is, a candidate ROI that is identified as an outlier based on one or more criteria (e.g., similarity to other candidate ROIs) may be discarded and may not be used to determine the final ROI 440. FIG. 4 further illustrates a particular example in which the final ROI 440 that is determined by intraframe ROI fusion 438 is used as an object bounding box 442 (e.g., for object tracking). For example, the object bounding box 442 may be an initial bounding box that is used to track the child as the child runs in the scene. However, it

will be appreciated that the final ROI 440 may be used for other computer vision (CV) applications (e.g., for object recognition, for reality augmentation, or scene analysis, among other alternatives).

[0038] Thus, FIG. 4 illustrates that the motion vector field 416 generated by a video encoder (e.g., video encoding hardware) may be used for segmentation and one-touch object selection. The example of one-touch object selection illustrated in FIG. 4 includes segmentation by region growing to generate multiple candidate ROIs and performing ROI fusion based on at least a subset of the candidate ROIs to determine a final ROI (e.g., for an object tracking application). While FIG. 4 illustrates a particular example that includes motion field generation 412, global/local motion estimation 414, segmentation by region grow 426, and intraframe ROI fusion 438, it will be appreciated that the order is not limiting. That is, alternative orders are possible, with more steps, fewer steps, different steps, concurrent steps, etc.

[0039] Referring to FIG. 5, a particular illustrative embodiment of a method of error handling in the context of motion-based segmentation for object selection is disclosed and generally designated 500. In the example illustrated in FIG. 5, error handling may be responsive to a one-touch input 502 (e.g., a user touch on a portion of an image that does not include an object with associated local motion). In FIG. 5, the one-touch input 502 is represented as a white dot on a patch of grass.

[0040] FIG. 5 illustrates that performing intraframe ROI fusion 504 responsive to a user touch on the grass may result in a final ROI 506 that exceeds a size threshold (e.g., for object tracking). In the context of object tracking, the size threshold may be based on an assumption that the user would not be tracking an object as large as the size of the final ROI 506. For tracking purposes, the size threshold may specify that the object be within a particular spatial range of the one-touch input 502. To illustrate, the size threshold may specify that the object be smaller than a maximum object size and larger than a minimum object size. Additionally or alternatively, the size threshold may specify a minimum aspect ratio and a maximum aspect ratio for an object.

[0041] In the particular embodiment illustrated in FIG. 5, error handling 508 may include generating a visual indication 510. The visual indication 510 may alert a user that the one-touch user input 502 was not successful in selecting the running child. The

visual indication 510 may prompt the user to provide another one-touch input. In some cases, the visual indication 510 may include a bounding box having a default size that is generated based on the X,Y coordinates of the one-touch user input 502.

[0042] While FIGS. 4 and 5 illustrate spatial segmentation for one-touch object selection, it will be appreciated that other types of segmentation may be used instead of or in addition to spatial segmentation. Further, while FIGS. 4 and 5 illustrate intraframe or spatial ROI fusion, it will be appreciated that other types of ROI fusion may be used instead or in addition to spatial ROI fusion. For example, FIG. 6 illustrates a particular illustrative embodiment of a method of color-based segmentation that includes temporal ROI fusion for one-touch object selection and generally designated 600. FIG. 6 illustrates that two-stage segmentation may be performed for multiple video frames to generate multiple candidate ROIs, and temporal ROI fusion may be used to generate the final ROI. In some embodiments, color-based segmentation may be performed when motion-based segmentation (e.g., as described with reference to FIGS. 4-5) fails.

[0043] FIG. 6 illustrates that the output of the color-based segmentation is a bounding box (as in the motion-based segmentation method described with respect to FIGS. 4-5), and the user input 602 is a one-touch user input (as in the motion-based segmentation method described with respect to FIGS. 4-5). By contrast, FIG. 6 illustrates a temporal dual-segmentation approach (e.g., a two stage segmentation approach) followed by temporal ROI fusion rather than spatial ROI fusion as described with respect to FIGS. 4-5. To illustrate, for color-based segmentation, a predetermined number of video frames may be identified for segmentation (e.g., five frames). Color-based segmentation may be performed for each of the five frames, and the method may include identifying consistent segmentation results among the five frames. That is, in the motion-based segmentation approached described for FIG. 4, the ROI fusion is done spatially, and in the particular example of color-based segmentation illustrated in FIG. 6, the ROI fusion may be done temporally.

[0044] In FIG. 6, a user input 602 may include a one-touch input. In response to the user input 602, a two-stage segmentation may be performed for multiple video frames. That is, processing of a particular video frame 604 may include a first stage segmentation 606 and a second stage segmentation 608, resulting in a candidate ROI

610 associated with the particular video frame 604. Multiple candidate ROIs may be generated, each associated with a particular video frame of the multiple video frames. In order to identify consistent segmentation results among the multiple video frames, temporal ROI fusion 612 may be performed to generate a final ROI 614.

[0045] For illustrative purposes only, FIG. 6 shows a first video frame 616 (“Frame N”), a second video frame 618 (“Frame N+1”), and a third video frame 620 (“Frame N+2”). However, it will be appreciated that color-based segmentation may be performed for an alternative number of frames. In FIG. 6, a user touch location 622 is shown in the first video frame 616. Due to camera motion or motion of objects in the scene, objects may move from frame to frame. FIG. 6 illustrates that the user touch location 622 may be propagated to subsequent frames. To illustrate, in the example of FIG. 6, the user touch location 622 is on the tip of the nose, and this point on the tip of the nose may be propagated from the first video frame 616 to the second video frame 618. Further, the user touch location 622 on the tip of the nose may be propagated from the second video frame 618 to the third video frame 620. In some cases, a motion vector field that is generated by a video encoder (as described above with respect to the motion vector field 416) may be used to propagate the user touch location 622 between frames.

[0046] For the first video frame 616, the user touch location 622 may be used to determine a starting region (e.g., a 5x5 box), and region growing may be used to grow the starting region into a mask. In some cases, if the mask fails to satisfy a size threshold (e.g., the mask is too large), region growing may be performed again using a larger starting region (e.g., a 7x7 box or a 9x9 box). In the color-based segmentation approach, region growing may be applied to red, green, and blue (RGB) color channel information, rather than X,Y coordinates (as in the motion-based approach of FIG. 4). Based on the mask, a first candidate ROI 624 may be generated.

[0047] FIG. 6 illustrates an example of segmentation using a seeded region grow method. That is, the user provides a seed in the form of a single touch point (i.e., the user touch location 622). In FIG. 6, a dual layer (also referred to herein as dual-stage) approach includes a first layer starting from a 5x5 box centered on the user touch location 622 that is grown into a region with area N (illustrated as the first stage

segmentation 606). In some cases, the area N may not satisfy a size threshold (e.g., the area N may be too small). Accordingly, a second layer starting from a box (centered on the user touch location 622) having a different size (e.g., an $M \times M$ box with M greater than 5 in this case) may be grown into a region with area R (illustrated as the second stage segmentation 608). In some cases, M may be determined based on N and may be proportional with N . In a particular embodiment, a maximum size may be determined based on $(1/3)\text{frameHeight} * (1/3)\text{frameWidth}$, while a minimum size may be 16×16 pixels (among other alternative sizes). Further, in some cases, there may be a maximum aspect ratio and a minimum aspect ratio threshold. To illustrate, the aspect ratio thresholds may exclude tall, thin boxes or flat, narrow boxes.

[0048] For the second video frame 618, the propagated user touch location 622 may determine another starting box (e.g., a 5×5 box), region growing using RGB color channel information may be used to grow the starting box into a mask, and a second candidate ROI 626 may be generated from the mask. Similarly, for the third video frame 620, the propagated user touch location 622 may determine another starting box (e.g., a 5×5 box), region growing using RGB color channel information may be used to grow the starting box into a mask, and a third candidate ROI 628 may be generated from the mask.

[0049] The temporal ROI fusion 612 may include determining the final ROI 614 based on at least a subset of the candidate ROIs. That is, at least a subset of the first candidate ROI 624, the second candidate ROI 626, and the third candidate ROI 628 may be used to determine the final ROI 614. FIG. 6 illustrates that the final ROI 614 may be used to generate an object bounding box 630 (e.g., for object tracking).

[0050] Referring to FIG. 7, a particular illustrative embodiment of a method of one-touch object selection by performing segmentation and ROI fusion is disclosed and generally designated 700. In an illustrative embodiment, the method 700 may be performed by the mobile device 102 of FIG. 1.

[0051] The method 700 includes receiving a user input (e.g., a one-touch input), at 702. For example, the user input may include the one-touch input 402 illustrated in FIG. 4 or the one-touch input 502 illustrated in FIG. 5. Alternatively, the user input may include

a non-touch input, such as gesture input, ultrasound sensor input corresponding to detection of a stylus or other device operated by the user, etc.

[0052] The method 700 includes performing segmentation to generate multiple candidate regions of interest (ROIs), at 704. For example, in some cases, segmentation may include the motion-based segmentation described with respect to FIG. 4. In other cases, the segmentation may include the color-based segmentation described with respect to FIG. 6. Alternatively, the segmentation may include both motion-based segmentation and color-based segmentation. To illustrate, both motion and color information may be examined when performing the segmentation. That is, both the XY coordinate information and the RGB color channel information may be used for segmentation.

[0053] The method 700 includes performing ROI fusion on at least a subset of the candidate ROIs to generate a final ROI, at 706. For example, performing ROI fusion may include performing the intraframe ROI fusion 438 described with respect to FIG. 4. As another example, performing ROI fusion may include performing the temporal ROI fusion 612 described with respect to FIG. 6.

[0054] Referring to FIG. 8, a particular illustrative embodiment of a method of object selection using a combination of motion-based and color-based segmentation along with ROI fusion is disclosed and generally designated 800. In an illustrative embodiment, the method 800 may be performed by the mobile device 102 of FIG. 1.

[0055] The method 800 includes receiving video frame(s), at 802, and performing motion-based segmentation to determine a motion ROI, at 804. In the context of motion-based segmentation, a sequence of video frames may be received in order to estimate motion. For example, referring to FIG. 4, the sequence of video frames 404 may be received, and the motion vector field 416 may be generated by a video encoder based on the sequence of video frames 404. As illustrated in FIG. 4, the segmentation by region growing 426 may include generating multiple candidate ROIs 428 and performing intraframe (spatial) ROI fusion 440 on at least a subset of the candidate ROIs 428.

[0056] At 806, the method 800 includes determining whether the ROI generated by the intraframe ROI fusion 440 represents a valid ROI. For example, as described above with respect to FIG. 4, in some cases the ROI generated by ROI fusion may not satisfy a size threshold. For example, in the context of object tracking, the size threshold may be based on an assumption that the user would not be tracking an object as large as the ROI determined based on intraframe ROI fusion. For tracking purposes, the size threshold may specify that the object be within a particular spatial range of the one-touch input 402. To illustrate, the size threshold may specify that the object be smaller than a maximum object size and larger than a minimum object size. Additionally or alternatively, the size threshold may specify a minimum aspect ratio and a maximum aspect ratio for an object.

[0057] When the motion ROI is valid, the method 800 may include generating a final ROI, at 816. That is, in the particular embodiment illustrated in FIG. 8, the fused ROI generated using the motion-based segmentation approach may be considered a higher priority or sufficient result, and the method 800 may not include performing color-based segmentation. In some cases, an object bounding box may be generated based on the final ROI. To illustrate, referring to FIG. 4, the object bounding box 442 may be generated based on the final ROI 440.

[0058] When the motion ROI is determined to be invalid at 806, the method 800 may include performing color-based segmentation to determine a color ROI for a particular video frame, at 808. That is, in the particular embodiment illustrated in FIG. 8, color-based segmentation may be performed when motion-based segmentation fails. To illustrate, referring to FIG. 6, color-based segmentation may be performed on the first video frame 616. For the first video frame 616, the user touch location 622 may be used to determine a starting region (e.g., a 5x5 box), and region growing may be used to grow the starting region into a mask. In some cases, if the mask is too large, region growing may be performed again using a larger starting region (e.g., a 7x7 box or a 9x9 box). In the color-based segmentation approach, region growing may be applied to red, green, and blue (RGB) color channel information, rather than X,Y coordinates (as in the motion-based segmentation approach of FIG. 4). Based on the mask, a first candidate ROI 624 may be generated.

[0059] The method 800 includes determining whether a particular (e.g., a maximum) frame number has been reached, at 810. That is, color-based segmentation may be performed for a particular number of frames (e.g., five frames), and the method 800 may return to 802 to receive information associated with another frame until the particular number of frames for color-based segmentation has been reached or until the motion ROI is valid. To illustrate, referring to the example of FIG. 6, three frames are illustrated. After performing the color-based segmentation on the first video frame 616 to determine the first candidate ROI 624, color-based segmentation may be performed on the second video frame 618 to determine the second candidate ROI 626. After performing the color-based segmentation on the second video frame 618 to determine the second candidate ROI 626, color-based segmentation may be performed on the third video frame 620.

[0060] When the particular number of frames has been reached at 810, the method 800 includes performing temporal ROI fusion of color ROIs, at 812. To illustrate, referring to FIG. 6, the temporal ROI fusion 612 may be performed on the first candidate ROI 624, the second candidate ROI 626, and the third candidate ROI 628. At 814, the method 800 includes determining whether the fused color ROI is valid. To illustrate, referring to FIG. 6, the final ROI 614 that represents the results of the temporal ROI fusion 612 of the candidate ROIs 624, 626, and 628 may be evaluated to determine validity. When the fused color ROI is valid, the method 800 proceeds to 816, where the final ROI resulting from the color-based segmentation (e.g., the final ROI 614 in FIG. 6) is determined to be the final ROI. In some cases, an object bounding box may be generated based on the final ROI. To illustrate, referring to FIG. 6, the object bounding box 630 may be generated based on the final ROI 614.

[0061] In particular embodiments, the method 700 of FIG. 7 and the method 800 of FIG. 8 may be implemented via hardware (e.g., a field-programmable gate array (FPGA) device, an application-specific integrated circuit (ASIC), etc.) of a processing unit, such as a central processing unit (CPU), a digital signal processor (DSP), or a controller, via a firmware device, or any combination thereof. As an example, the method 700 of FIG. 7 and the method 800 of FIG. 8 can be performed by a processor that executes instructions, as described with respect to FIG. 9.

[0062] Referring to FIG. 9, a block diagram of a particular illustrative embodiment of an electronic device including an object selection module 902 (e.g., a one-touch object selection module) is depicted and generally designated 900. The device 900 includes a processor 910, such as a central processing unit (CPU), coupled to a memory 932 and also coupled to camera controller 982. The camera controller 982 is coupled to a camera 980. In the example of FIG. 9, the object selection module 902 is shown as instructions within the memory 932, and these instructions can be executed by the processor 910 to perform all or a portion of one or more methods described herein (e.g., the method 700 of FIG. 7 and the method 800 of FIG. 8). In alternative embodiments, all or a portion of the object selection module 902 could be implemented using hardware (e.g., within the processor 910).

[0063] The processor 910 may include a video encoder 904 configured to execute a motion estimation stage 906 and one or more other stages 908. In an illustrative example, the camera 980 includes the camera 104 of FIG. 1. Alternatively, the video encoder 904 may be instructions stored in the memory 932 and executed by the processor 910. In some cases, execution of the motion estimation stage 906 may result in generation of a motion vector field (e.g., the motion vector field 416 of FIG. 4). Further, in some cases, the other stage(s) 908 may be turned off or disabled, as the other stage(s) may not be used for one-touch object selection.

[0064] FIG. 9 also shows a display controller 926 that is coupled to the processor 910 and to a display 928. The display controller 926 may correspond to the output interface 106 depicted in FIG. 1. A coder/decoder (CODEC) 934 can also be coupled to the processor 910. A speaker 936 and a microphone 938 can be coupled to the CODEC 934.

[0065] FIG. 9 also indicates that a wireless controller 940 can be coupled to the processor 910 and to an antenna 942. The wireless controller 940 may correspond to the network interface 116 depicted in FIG. 1. In a particular embodiment, the processor 910, the display controller 926, the memory 932, the CODEC 934, the wireless controller 940, and the camera controller 982 are included in a system-in-package or system-on-chip device 922. In a particular embodiment, an input device 930 and a power supply 944 are coupled to the system-on-chip device 922. The input device 930

may correspond to the input interface 108 of FIG. 1. Moreover, in a particular embodiment, as illustrated in FIG. 9, the display 928, the input device 930, the speaker 936, the microphone 938, the camera 980, the antenna 942, and the power supply 944 are external to the system-on-chip device 922. However, each of the display 928, the input device 930, the speaker 936, the microphone 938, the camera 980, the antenna 942, and the power supply 944 can be coupled to a component of the system-on-chip device 922, such as an interface or a controller.

[0066] In conjunction with the described embodiments, an apparatus is disclosed that includes means for receiving a one-touch user input. The one-touch user input may be associated with X,Y coordinates of a first image of a sequence of images. The means for receiving may include the input interface 108, the camera 104, the input device 930, the camera 980, or any combination thereof. The means for receiving may also include the input device 930 of FIG. 9, one or more other devices or circuits configured to receive data associated with the one-touch user input (e.g., a touchscreen of a mobile phone), or any combination thereof.

[0067] The apparatus further includes means for generating a motion vector field responsive to the one-touch user input. The means for generating may include the object selection module 122, the processor 910, the video encoder 904, one or more other devices or circuits configured to generate a motion vector field, or any combination thereof.

[0068] The apparatus further includes means for generating multiple candidate regions of interest (ROIs) by segmentation. The means for generating the multiple candidate ROIs may include the object selection module 122, the processor 910, the video encoder 904, one or more other devices or circuits configured to determine a region of interest by segmentation, or any combination thereof. The apparatus further includes means for generating a final ROI based on at least a subset of the candidate ROIs. The means for generating the final ROI may include the processor 910, one or more other devices or circuits configured to analyze each of the candidate ROIs generated by segmentation, or any combination thereof.

[0069] The apparatus may further include means for displaying an object bounding box associated with the final ROI. For example, the means for displaying may include the

output interface 106, the display controller 926, the display 928, or any combination thereof.

[0070] Those of skill would further appreciate that the various illustrative logical blocks, configurations, modules, circuits, and algorithm steps described in connection with the embodiments disclosed herein may be implemented as electronic hardware, computer software executed by a processor, or combinations of both. Various illustrative components, blocks, configurations, modules, circuits, and steps have been described above generally in terms of their functionality. Whether such functionality is implemented as hardware or processor executable instructions depends upon the particular application and design constraints imposed on the overall system. Skilled artisans may implement the described functionality in varying ways for each particular application, but such implementation decisions should not be interpreted as causing a departure from the scope of the present disclosure.

[0071] The steps of a method or algorithm described in connection with the embodiments disclosed herein may be embodied directly in hardware, in a software module executed by a processor, or in a combination of the two. A software module may reside in random access memory (RAM), flash memory, read-only memory (ROM), programmable read-only memory (PROM), erasable programmable read-only memory (EPROM), electrically erasable programmable read-only memory (EEPROM), registers, hard disk, a removable disk, a compact disc read-only memory (CD-ROM), or any other form of storage medium known in the art. An exemplary non-transitory (e.g., tangible) storage medium is coupled to the processor such that the processor can read information from, and write information to, the storage medium. In the alternative, the storage medium may be integral to the processor. The processor and the storage medium may reside in an application-specific integrated circuit (ASIC). The ASIC may reside in a computing device or a user terminal. In the alternative, the processor and the storage medium may reside as discrete components in a computing device or user terminal.

[0072] The previous description of the disclosed embodiments is provided to enable a person skilled in the art to make or use the disclosed embodiments. Various modifications to these embodiments will be readily apparent to those skilled in the art,

and the principles defined herein may be applied to other embodiments without departing from the scope of the disclosure. Thus, the present disclosure is not intended to be limited to the embodiments shown herein but is to be accorded the widest scope possible consistent with the principles and novel features as defined by the following claims.

WHAT IS CLAIMED IS:

1. A method comprising:
receiving user input;
in response to the user input, performing segmentation at a processor to generate multiple candidate regions of interest; and
performing region of interest fusion at the processor to generate a final region of interest.
2. The method of claim 1, wherein the user input includes a one-touch user input received at a touch screen.
3. The method of claim 1, wherein the user input includes a gesture.
4. The method of claim 1, wherein the user input is detected by an ultrasound sensor.
5. The method of claim 1, wherein the user input is associated with a computer vision based automatic object detection application.
6. The method of claim 1, further comprising:
generating an object bounding box based on the final region of interest; and
displaying the object bounding box.
7. The method of claim 6, wherein an object of interest corresponding to the user input is substantially included within the object bounding box.
8. The method of claim 1, wherein the segmentation includes motion-based segmentation of image data.
9. The method of claim 1, wherein the segmentation includes color-based segmentation of image data.

10. The method of claim 1, wherein the segmentation includes a combination of motion-based segmentation and color-based segmentation.

11. The method of claim 10, wherein the segmentation includes:
receiving multiple video frames;
performing motion-based segmentation based on the multiple video frames to generate a first motion region of interest;
determining whether the first motion region of interest is valid; and
in response to determining that the first motion region of interest is invalid, performing color-based segmentation on a first video frame of the multiple video frames to determine a first color region of interest.

12. The method of claim 11, further comprising generating the final region of interest based on the first motion region of interest in response to determining that the first motion region of interest is valid.

13. The method of claim 11, further comprising:
performing color-based segmentation on a second video frame of the multiple video frames to determine a second color region of interest;
performing temporal fusion of the first color region of interest and the second color region of interest to generate a temporally fused color region of interest;
determining whether the temporally fused color region of interest is valid; and
in response to determining that the temporally fused color region of interest is valid, generating the final region of interest based on the temporally fused color region of interest.

14. The method of claim 13, further comprising performing an error handling operation in response to determining that the temporally fused color region of interest is invalid.

15. The method of claim 14, wherein performing the error handling operation includes displaying a visual indication.

16. An apparatus comprising:
 - a processor;
 - an input device configured to receive user input;
 - a video encoder configured to generate a motion vector field responsive to the user input;
 - a segmentation component executable by the processor to perform segmentation to generate multiple candidate regions of interest responsive to receiving the motion vector field from the video encoder; and
 - a fusion component executable by the processor to perform region of interest fusion to generate a final region of interest.
17. The apparatus of claim 16, wherein the motion vector field is generated based on a sequence of video frames corresponding to a video stream.
18. The apparatus of claim 16, wherein the video encoder is configured to generate the motion vector field by executing a subset of video encoding stages associated with motion estimation.
19. The apparatus of claim 18, wherein during generation of the motion vector field, the video encoder is configured to not execute a second subset of video encoding stages that are not associated with motion estimation.
20. The apparatus of claim 16, wherein the user input is associated with a first set of X,Y coordinates in the motion vector field.
21. The apparatus of claim 20, wherein the first set of X,Y coordinates are used as a first seed for region growing to generate a first candidate region of interest of the multiple candidate regions of interest.
22. The apparatus of claim 21, wherein a second set of X,Y coordinates that represent neighboring X,Y coordinates with respect to the first set of X,Y coordinates are used as a second seed for region growing to generate a second candidate region of interest of the multiple candidate regions of interest.

23. The apparatus of claim 22, wherein the second set of X,Y coordinates are offset by a particular number of pixels with respect to the first set of X,Y coordinates.

24. The apparatus of claim 23, wherein the particular number of pixels is an integer value that represents an offset in at least one of a positive X direction, a negative X direction, a positive Y direction, or a negative Y direction.

25. An apparatus comprising:

means for receiving a user input;

means for generating a motion vector field responsive to the user input;

means for performing segmentation to generate multiple candidate regions of interest; and

means for performing region of interest fusion to generate a final region of interest.

26. A computer-readable storage device storing instructions that, when executed by a processor, cause the processor to perform operations comprising:

receiving a user input;

in response to the user input, performing segmentation to generate multiple candidate regions of interest; and

performing region of interest fusion to generate a final region of interest.

27. The computer-readable storage device of claim 26, wherein the operations further comprise performing an error handling operation in response to determining that the final region of interest is invalid.

28. The computer-readable storage device of claim 27, wherein the error handling operation includes generating a display that includes a visual indication of an object bounding box generated based on the final region of interest.

29. The computer-readable storage device of claim 28, wherein the visual indication prompts a user to provide a second user input.

30. The computer-readable storage device of claim 27, wherein the final region of interest is determined to be invalid in response to determining that the final region of interest does not satisfy a size threshold.

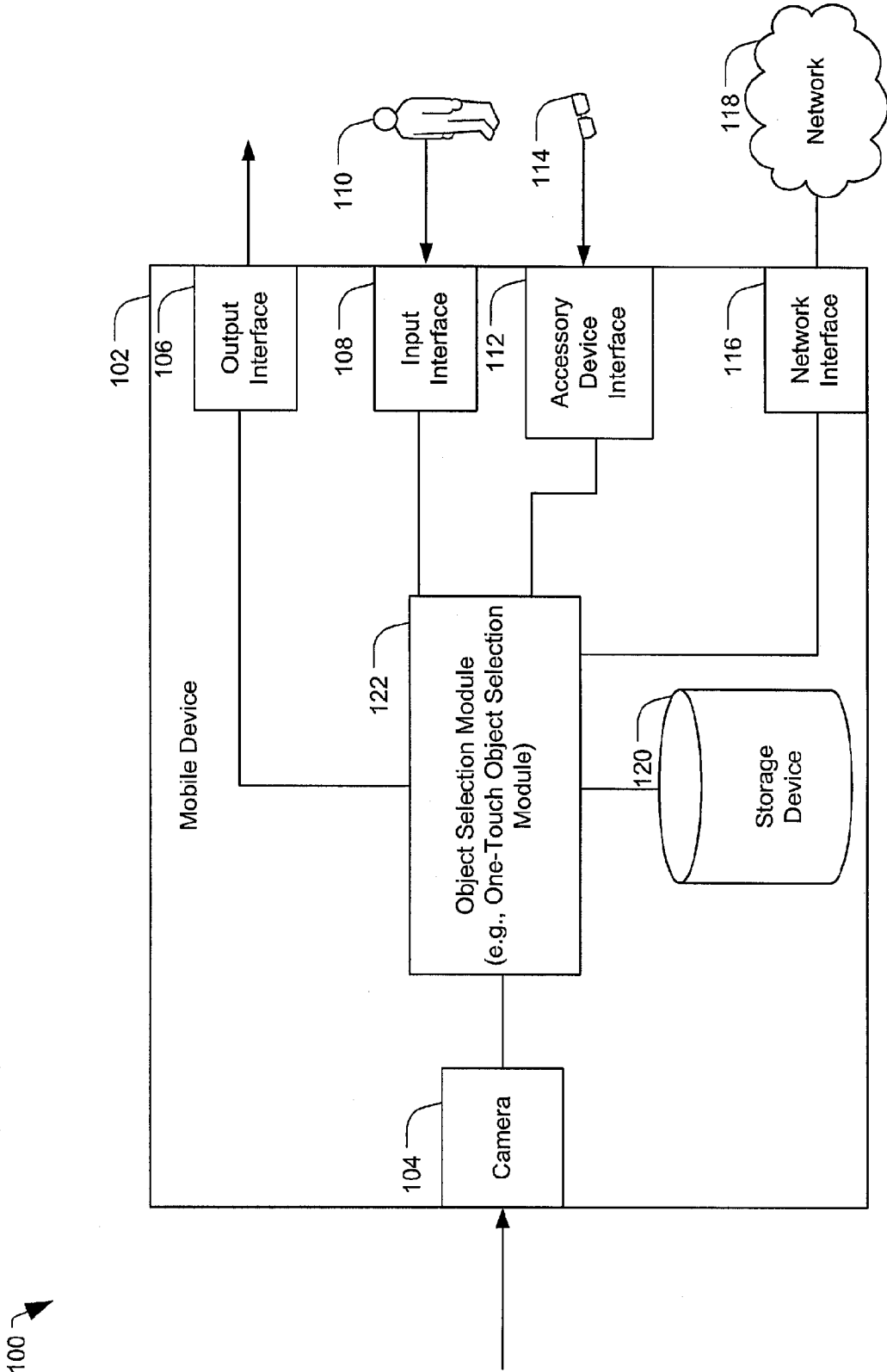


FIG. 1

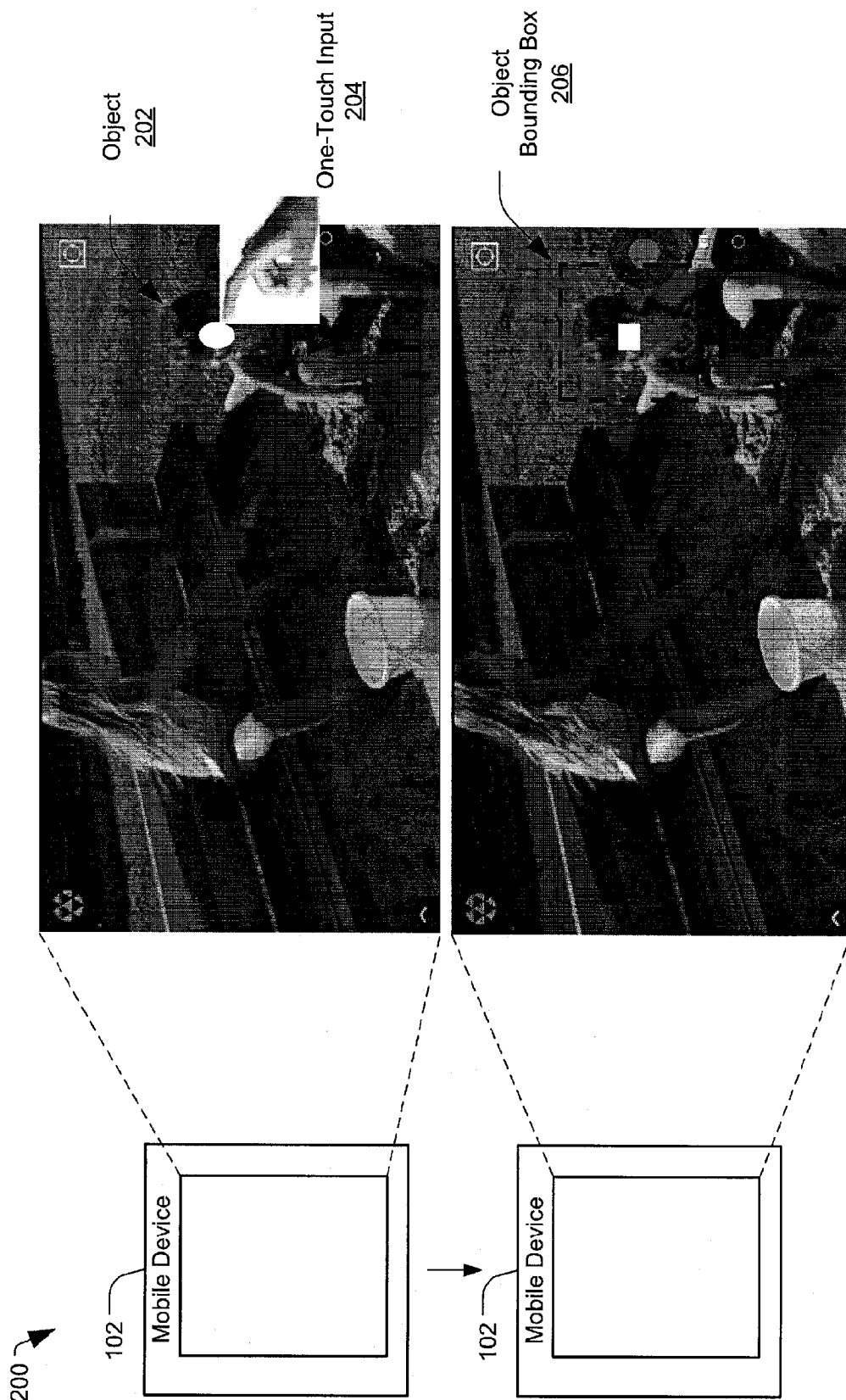


FIG. 2

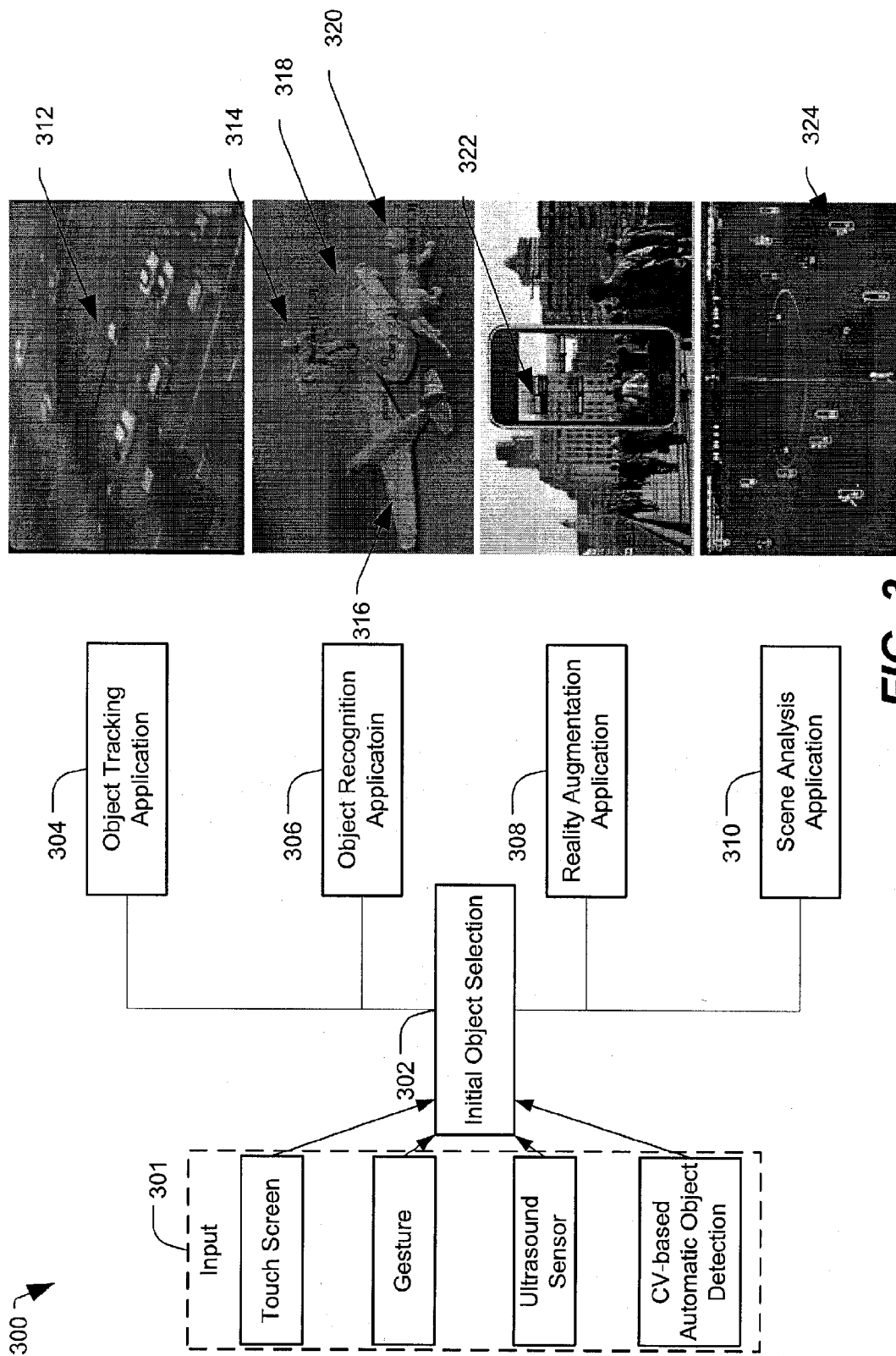


FIG. 3

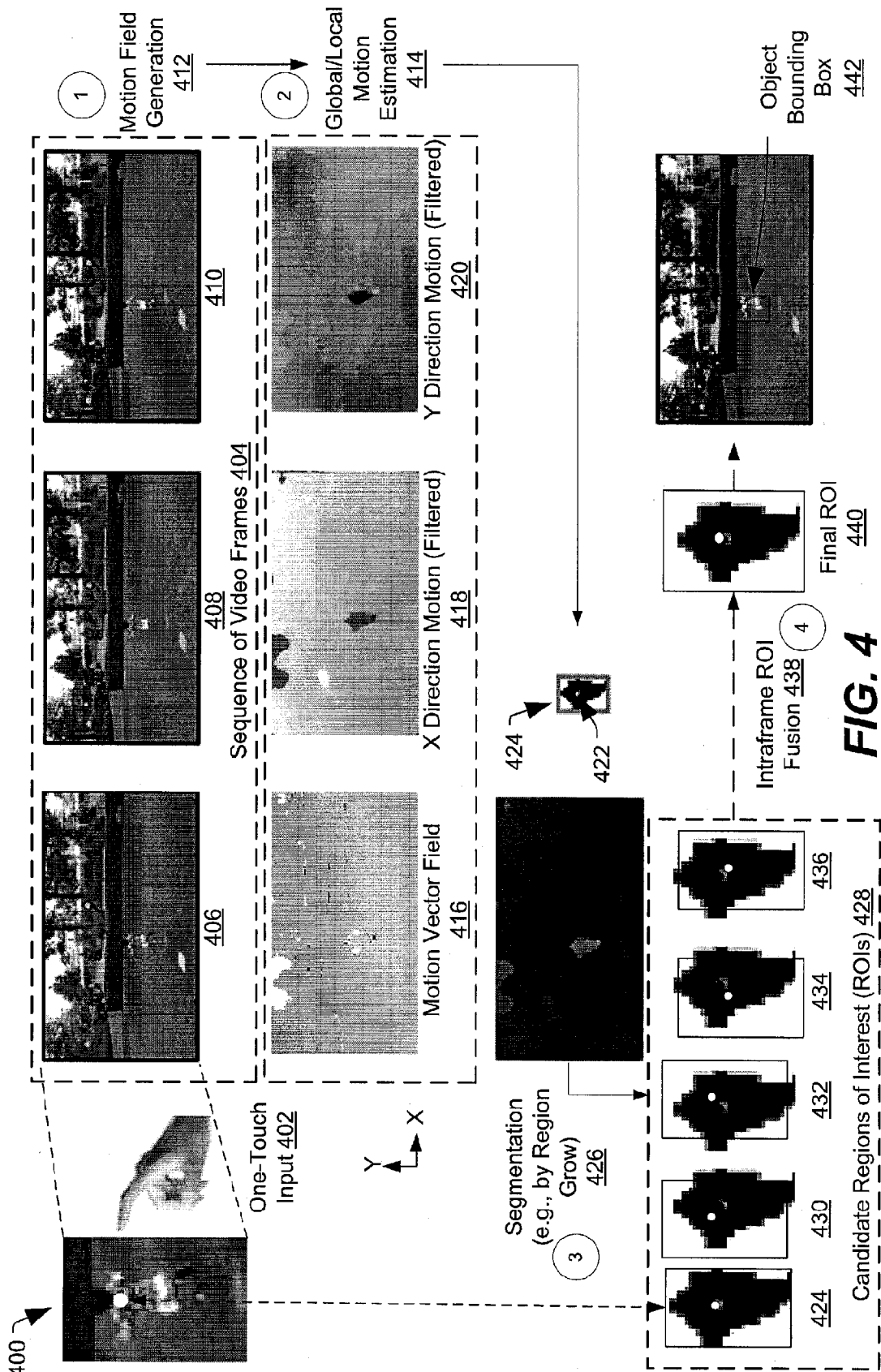


FIG. 4

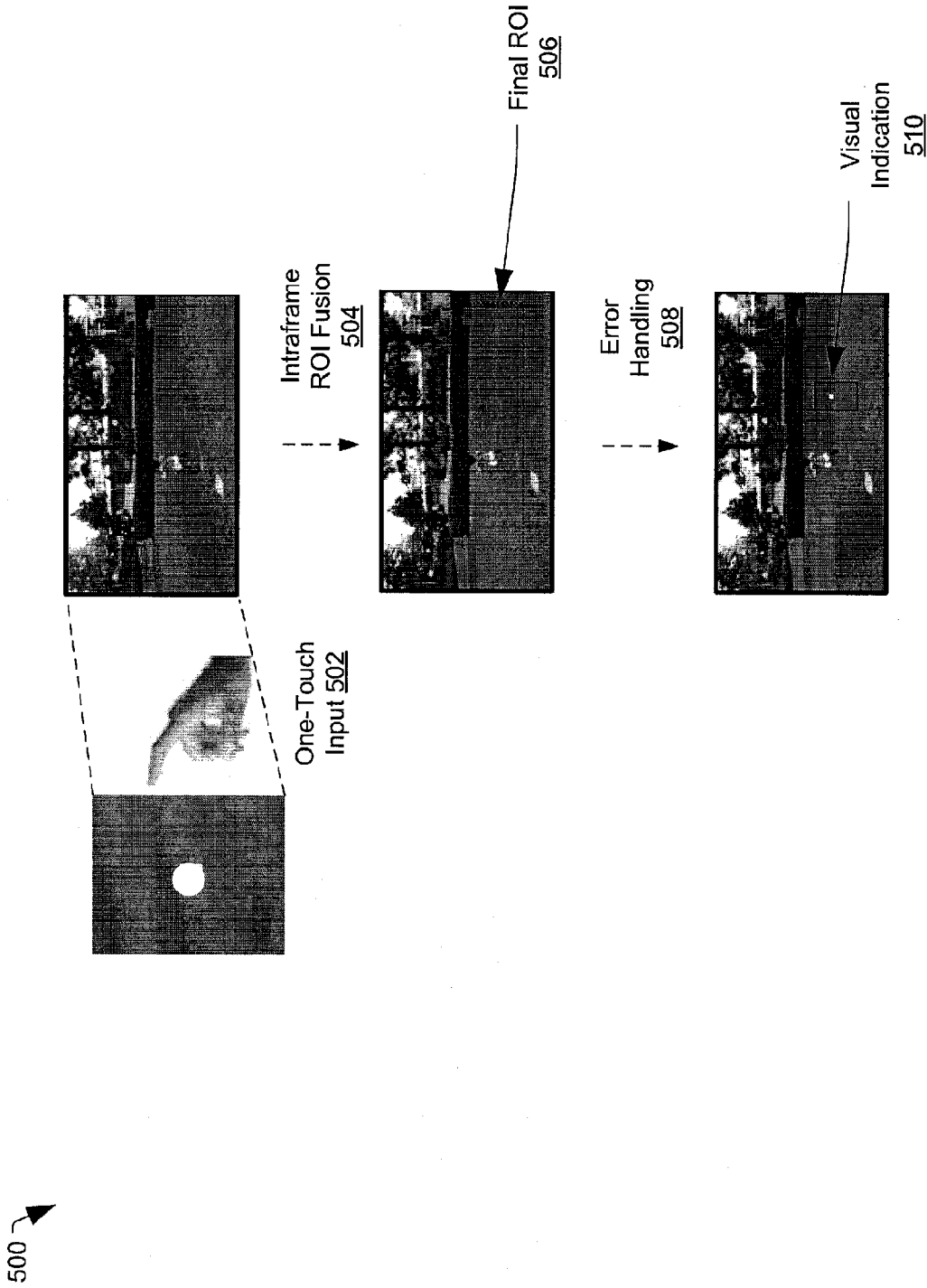


FIG. 5

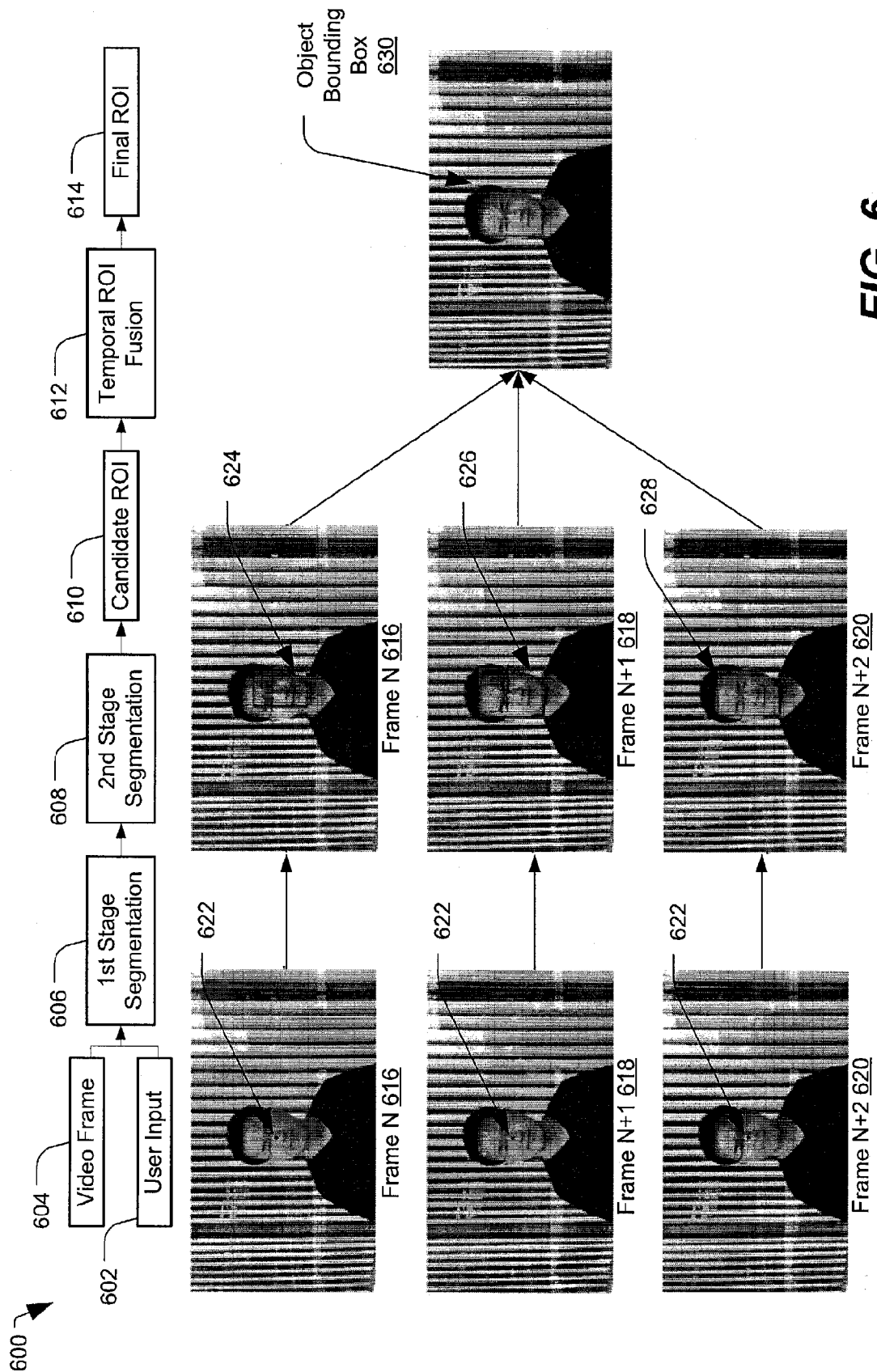
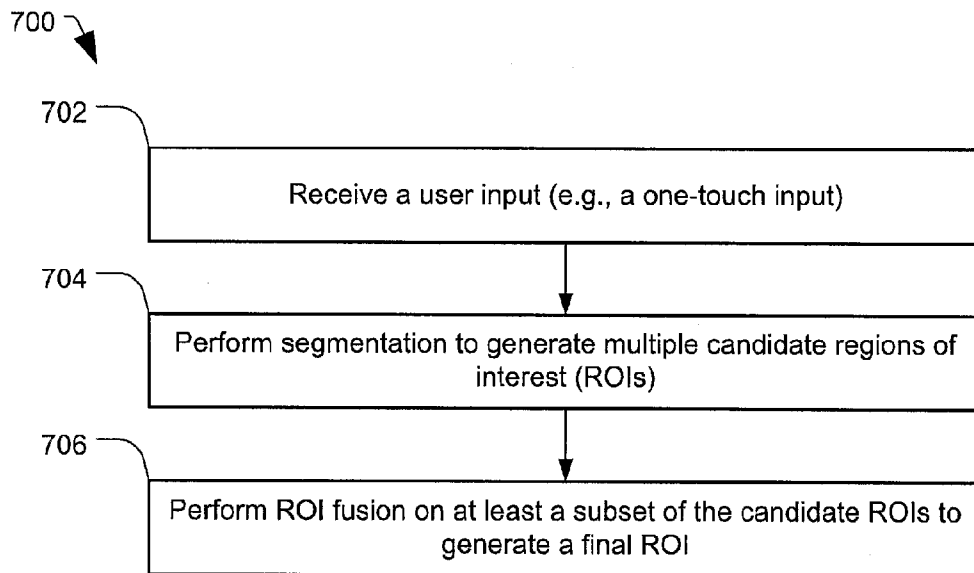
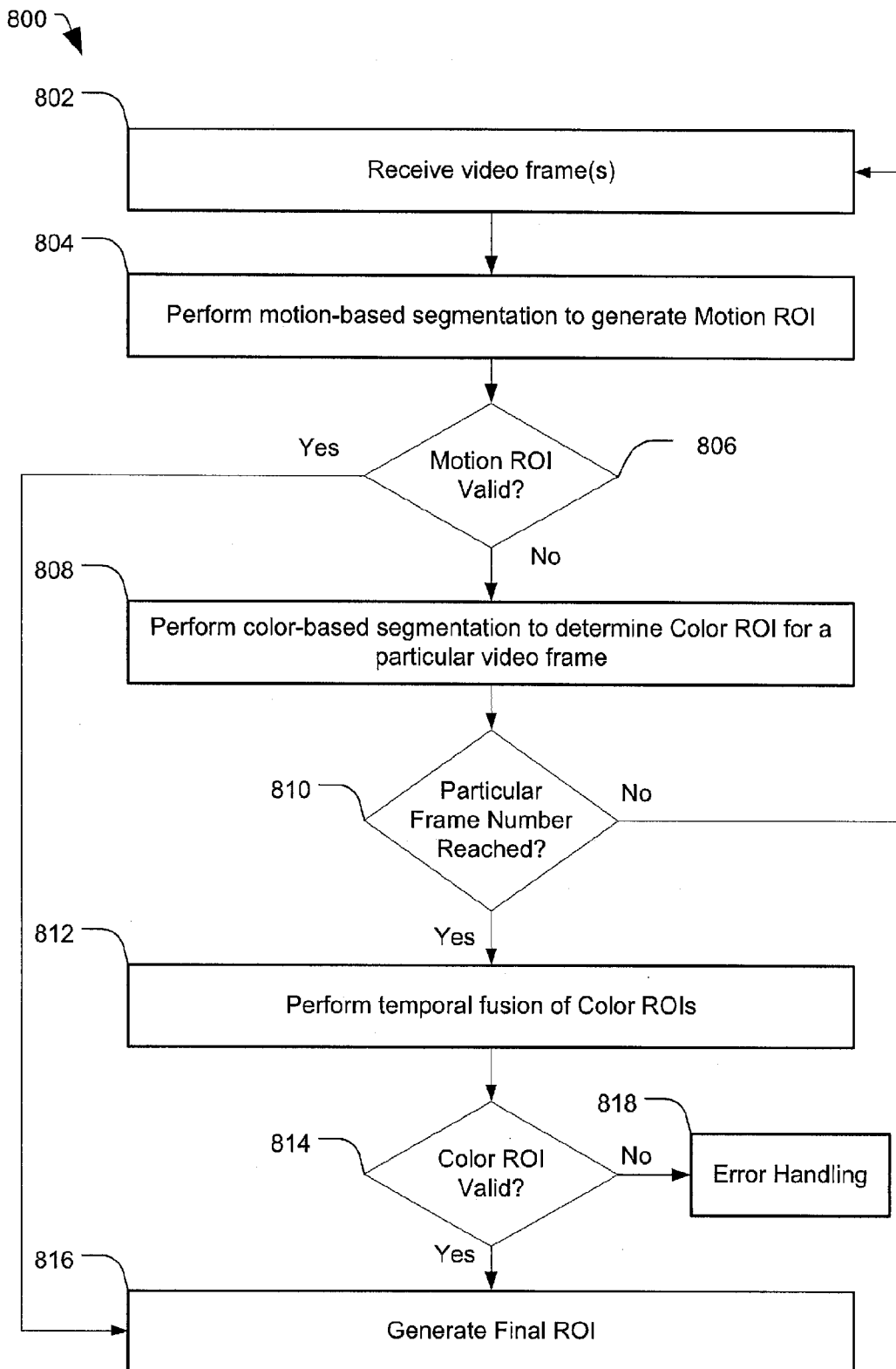
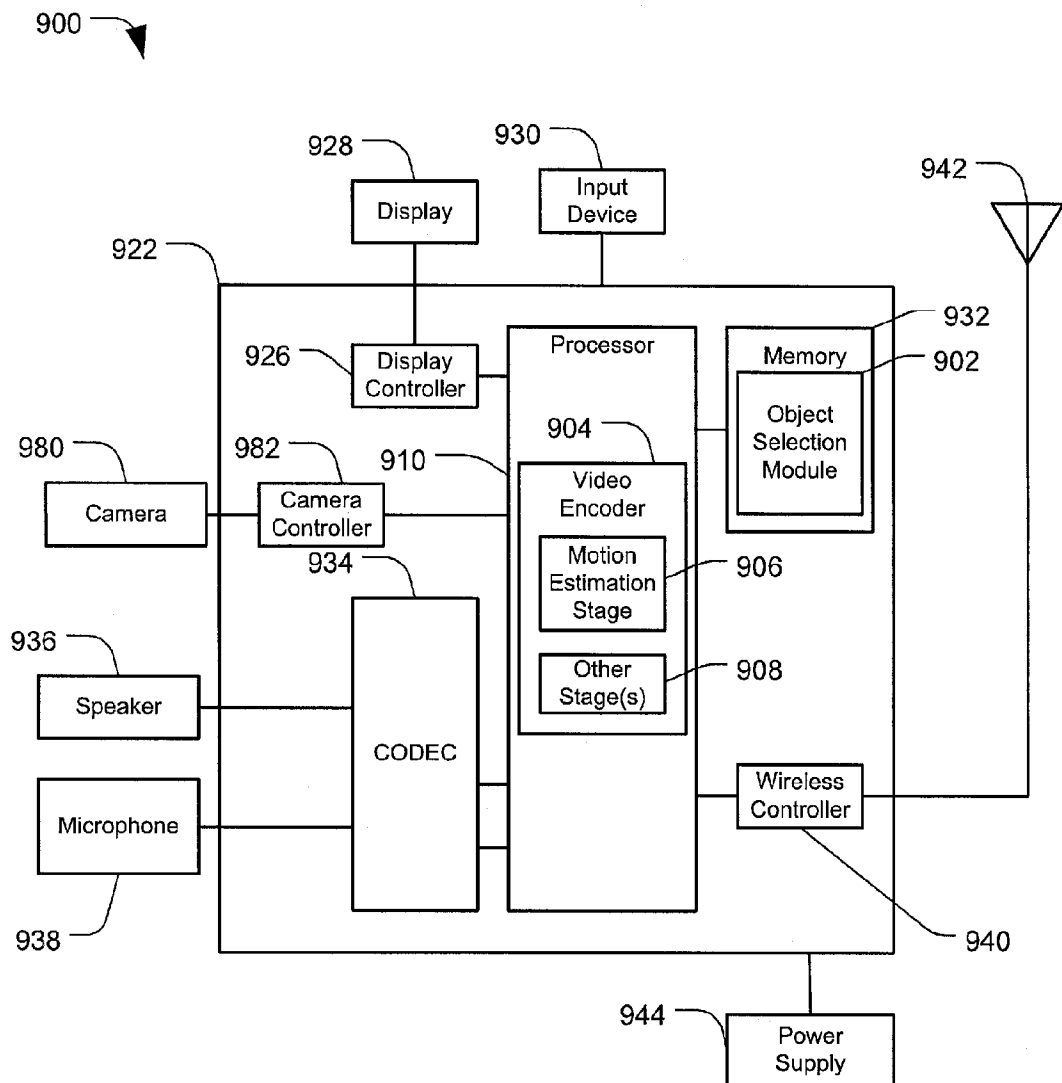


FIG. 6

**FIG. 7**

**FIG. 8**

**FIG. 9**

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2014/085381

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/34(2006.01)i; G06T 1/00(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K; G06T

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI, CNPAT, EPODOC, WPI: region of interest, ROI, segmentation, segment, divide, division, dismember, segmentation, fusion, combine, assemble, compound, video, input, touch, bound box

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2007183661 A1 (EL-MALEH, KHALED HELMI) 09 August 2007 (2007-08-09) description, paragraphs [0093]-[0099] and figures 14-18	1-30
X	US 2011158558 A1 (NOKIA CORPORATION) 30 June 2011 (2011-06-30) description, paragraphs [0051]-[0062] and figures 3b-5a	1-30
A	CN 103366359 A (SONY CORPORATION) 23 October 2013 (2013-10-23) the whole document	1-30
A	CN 102682454 A (UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA) 19 September 2012 (2012-09-19) the whole document	1-30

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance
 “E” earlier application or patent but published on or after the international filing date
 “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
 “O” document referring to an oral disclosure, use, exhibition or other means
 “P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
 “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
 “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
 “&” document member of the same patent family

Date of the actual completion of the international search

14 April 2015

Date of mailing of the international search report

29 April 2015

Name and mailing address of the ISA/CN

**STATE INTELLECTUAL PROPERTY OFFICE OF THE
P.R.CHINA(ISA/CN)
6,Xitucheng Rd., Jimen Bridge, Haidian District, Beijing
100088, China**

Authorized officer

WEI,Ling

Facsimile No. (86-10)62019451

Telephone No. (86-10)61648264

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No. PCT/CN2014/085381

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2007183661	A1	09 August 2007	JP	2009526331	A	16 July 2009
				AT	520102	T	15 August 2011
				CN	101375312	A	25 February 2009
				EP	1984896	A1	29 October 2008
				US	2012189168	A1	26 July 2012
				KR	20080100242	A	14 November 2008
				EP	2381420	A1	26 October 2011
				EP	2378486	A1	19 October 2011
				WO	2007092906	A1	16 August 2007
				INMUMNP200801247		E	12 September 2008
				US	2011158558	A1	30 June 2011
AU	2009357597	A1	05 July 2012				
KR	20120109591	A	08 October 2012				
CA	2785746	A1	07 July 2011				
RU	2012132016	A	10 February 2014				
WO	2011079442	A1	07 July 2011				
CN	102687140	A	19 September 2012				
INC HENP201206681		E	29 November 2013				
CN	103366359	A	23 October 2013	JP	2013214234	A	17 October 2013
				US	2013259400	A1	03 October 2013
CN	102682454	A	19 September 2012	None			