

CLAIMS

1. A method for detecting objects in video data, comprising:
  - determining object-label probability values for spatial elements of frames of
  - 5 video data using a two-dimensional image classifier, wherein the object-label probability value for each of the spatial elements indicates a probability that the respective spatial element is an observation of a particular object;
  - identifying surface elements in a three-dimensional surface element representation of a space observed in the frames of video data that correspond to the
  - 10 spatial elements, wherein a correspondence between a spatial element and a surface element is determined based on a projection of the surface element representation using an estimated pose for a frame; and
  - updating object-label probability values for the surface elements based on the object-label probability values for corresponding spatial elements to provide a
  - 15 semantically-labelled three-dimensional surface element representation of objects present in the video data, wherein the object-label probability value for each of the surface elements indicates a probability that the respective surface element represents the particular object.
- 20 2. The method of claim 1, wherein, during processing of said video data, the method comprises:
  - detecting a loop closure event and applying a spatial deformation to the surface element representation, the spatial deformation modifying three-dimensional positions of surface elements in the surface element representation,
  - 25 wherein the spatial deformation modifies the correspondence between spatial elements and surface elements of the surface element representation such that, after the spatial deformation, object-label probability values for a first surface element are updated using object-label probability values for spatial elements that previously corresponded to a second surface element.
- 30 3. The method of claim 1 or claim 2, comprising:

processing the frames of video data without a pose graph to generate the three-dimensional surface element representation, including, on a frame-by-frame basis:

5                   comparing a rendered frame generated using the three-dimensional surface element representation with a video data frame from the frames of video data to determine a pose of a capture device for the video data frame; and  
                    updating the three-dimensional surface element representation using the pose and image data from the video data frame.

4.           The method of claim 3, wherein:  
10           a subset of the frames of video data used to generate the three-dimensional surface element representation are input to the two-dimensional image classifier.

5.           The method of any one of the preceding claims, wherein the frames of video data comprise at least one of colour data, depth data and normal data; and  
15           wherein the two-dimensional image classifier is configured to compute object-label probability values based on at least one of colour data, depth data and normal data for a frame.

6.           The method of any one of the preceding claims, wherein the two-dimensional  
20 image classifier comprises a convolutional neural network.

7.           The method of claim 6, wherein the convolutional neural network is configured to output the object-label probability values as a set of pixel maps for each frame of video data, each pixel map in the set corresponding to a different object label in a set of  
25 available object labels.

8.           The method of claim 6 or claim 7, wherein the two-dimensional image classifier comprises a deconvolutional neural network communicatively coupled to the output of the convolutional neural network.

9. The method of any one of the preceding claims, comprising, after the updating of the object-label probability values for the surface elements:  
regularising the object-label probability values for the surface elements.
- 5 10. The method of claim 9, wherein regularising comprises:  
applying a conditional random field to the object-label probability values for surface elements in the surface element representation.
- 10 11. The method of claim 9 or claim 10, wherein regularising the object-label probability values comprises:  
regularising the object-label probability values assigned to surface elements based on one or more of: surface element positions, surface element colours, and surface element normals.
- 15 12. The method of any one of the preceding claims, comprising:  
replacing a set of one or more surface elements with a three-dimensional object definition based on the object-label probability values assigned to said surface elements.
- 20 13. The method of any one of the preceding claims, comprising:  
annotating surface elements of a three-dimensional surface element representation of a space with object-labels to provide an annotated representation;  
generating annotated frames of video data from the annotated representation based on a projection of the annotated representation, the projection using an estimated pose for each annotated frame, each annotated frame comprising spatial elements with  
25 assigned object-labels; and  
training the two-dimensional image classifier using the annotated frames of video data.
- 30 14. The method of any one of the preceding claims, comprising:  
obtaining a first frame of video data corresponding to an observation of a first portion of an object;

generating an image map for the first frame of video data using the two-dimensional image classifier, said image map indicating the presence of the first portion of the object in an area of the first frame; and

5 determining that a surface element does not project onto the area in the first frame and as such not updating object-label probability values for the surface element based image map values in said area;

wherein following detection of a loop closure event the method comprises:

modifying a three-dimensional position of the surface element;

10 obtaining a second frame of video data corresponding to a repeated observation of the first portion of the object;

generating an image map for the second frame of video data using the two-dimensional image classifier, said image map indicating the presence of the first portion of the object in an area of the second frame;

15 determining that the modified first surface element does project onto the area of the second frame following the loop closure event; and

updating object-label probability values for the surface element based on the image map for the second frame of video data,

wherein the object-label probability values for the surface element include fused object predictions for the surface element from multiple viewpoints.

20

15. Apparatus for detecting objects in video data comprising:

25 an image-classifier interface to receive two-dimensional object-label probability distributions for spatial elements of individual frames of video data, wherein the object-label probability distribution for each of the spatial elements includes a set of object-label probability values, and each of the set of object-label probability values indicates a probability that the respective spatial element is an observation of a different respective object;

30 a correspondence interface to receive data indicating, for a given frame of video data, a correspondence between spatial elements within the given frame and surface elements in a three-dimensional surface element representation, said correspondence

being determined based on a projection of the surface element representation using an estimated pose for the given frame; and

a semantic augmenter to iteratively update object-label probability values assigned to individual surface elements in the three-dimensional surface element representation,

wherein the semantic augmenter is configured to use, for a given frame of video data, the data received by the correspondence interface to apply the two-dimensional object-label probability distributions received by the image classifier interface to a set of object-label probability values assigned to corresponding surface elements, wherein each of the set of object-label probability values for each of the surface elements indicates a probability that the respective surface element represents a different respective object.

16. The apparatus of claim 15, wherein:

the correspondence interface is configured to provide an updated correspondence following a spatial deformation of the surface element representation, the spatial deformation enacting a loop closure within the video data, and

the semantic augmenter is configured to use the updated correspondence to update object-label probability values for a first surface element using object-label probability values for spatial elements that previously corresponded to a second surface element.

17. The apparatus of claim 15 or claim 16, wherein the image-classifier interface is configured to receive a plurality of image maps corresponding to a respective plurality of object labels for a given frame of video data, each image map having pixel values indicative of probability values for an associated object label.

18. The apparatus of any one of claims 15 to 17, comprising:

a regulariser to regularise the object-label probability values assigned to the surface elements of the surface element representation.

19. The apparatus of claim 18, wherein the regulariser is configured to apply a conditional random field to the object-label probability values for surface elements in the surface element representation.
- 5 20. The apparatus of claim 18 or claim 19, wherein the regulariser is configured to regularise the object-label probability values assigned to surface elements based on one or more of: surface element positions, surface element colours, and surface element normals.
- 10 21. The apparatus of any one of claims 15 to 20, wherein the semantic augments is configured to replace a set of one or more surface elements with a three-dimensional object definition based on the object-label probability values assigned to said surface elements.
- 15 22. The apparatus of any one of claims 15 to 21, wherein each surface element in the surface element representation comprises at least data defining a position of the surface element in three-dimensions and data defining a normal vector for the surface element in three-dimensions, and wherein each surface element represents a two-dimensional area in three-dimensional space.
- 20 23. A video processing system for detecting objects present in video data comprising:  
the apparatus of any one of claims 15 to 22;  
a video acquisition interface to obtain frames of video data from a capture  
25 device, said frames of video data resulting from relative movement between the capture device and a three-dimensional space over time; and  
a simultaneous location and mapping (SLAM) system communicatively coupled to the correspondence interface of the apparatus to generate a surface element representation of the three-dimensional space based on the obtained frames of video  
30 data,

wherein the SLAM system is configured to apply a spatial deformation to the surface element representation to close loops of observation within the frames of video data, said spatial deformation resulting in a new three-dimensional position for at least one modified surface element in the surface element representation.

5

24. The video processing system of claim 23, wherein the SLAM system comprises:  
a segmenter configured to segment the three-dimensional surface element representation into at least active and inactive portions based on at least one representation property,

10

wherein the SLAM system is configured to compute an active rendered frame based on a projection from the active portions of the surface element representation to update said representation over time; and

a registration engine configured to align active portions of the three-dimensional surface element representation with inactive portions of the three-dimensional surface element representation over time, the registration engine being configured to:

15

compute an inactive rendered frame based on a projection from the inactive portions of the three-dimensional surface element representation;

determine a spatial deformation that aligns the active rendered frame with the inactive rendered frame; and

20

update the three-dimensional surface element representation by applying the spatial deformation.

25. The video processing system of claim 23 or claim 24, wherein the SLAM system comprises:

25

a frame-to-model tracking component configured to compare the active rendered frame to a provided frame from said video data to determine an alignment of the active portions of the three-dimensional surface element representation with the video data.

30

26. The video processing system of any one of claims 24 to 25, wherein the registration engine is configured to use a deformation graph to align active portions of

the three-dimensional surface element representation with inactive portions of the three-dimensional surface element representation, the deformation graph being computed based on an initialisation time for surface elements, the deformation graph indicating a set of surface-element neighbours for a given surface element that are to be used to  
 5 modify the given surface element during alignment.

27. The video processing system of any one of claims 23 to 26, comprising:  
 a two-dimensional image classifier communicatively coupled to the image-classifier interface to compute object-label probability distributions for frames of the  
 10 video data obtained from the video acquisition interface.

28. The video processing system of claim 27, wherein the two-dimensional image classifier comprises a convolutional neural network.

15 29. The video processing system of claim 28, wherein the convolutional neural network is configured to output the object-label probability values as a set of pixel maps for each frame of video data.

20 30. The video processing system of claim 28 or claim 29, wherein the two-dimensional image classifier comprises a deconvolutional neural network communicatively coupled to the output of the convolutional neural network.

31. A robotic device comprising:  
 at least one capture device to provide frames of video data comprising one or  
 25 more of depth data and colour data, said depth data indicating a distance from the capture device for a plurality of image elements;  
 the apparatus of any one of claims 15 to 22 or the video processing system of any one of claims 23 to 30;  
 one or more movement actuators to move the robotic device with the three-  
 30 dimensional space; and  
 a navigation engine to control the one or more movement actuators,

wherein the navigation engine is configured to access the object-label probability values assigned to individual surface elements in the three-dimensional surface element representation to navigate the robotic device within the three-dimensional space.

5

32. The robotic device of claim 31, wherein the navigation engine is configured to identify entry and exit points for a room based on the object-label probability values assigned to surface elements in the three-dimensional surface element representation.

10

33. A mobile computing device comprising:

at least one capture device arranged to record frames of video data comprising one or more of depth data and colour data, said depth data indicating a distance from the capture device for a plurality of image elements, and

15

the apparatus of any one of claims 15 to 22 or the video processing system of any one of claims 23 to 30.

34. A non-transitory computer-readable storage medium comprising computer-executable instructions which, when executed by a processor, cause a computing device to perform the video processing method of any one of claims 1 to 14.