Claims:

1. A method for manipulating an initial video stream captured by a camera in a video conferencing endpoint into multiple views corresponding to regions of interest, the method comprising:

detecting objects of one or more predefined types in a frame of the initial video stream;

selecting a plurality of crop regions from the frame of the initial video stream, each crop region including at least one bounding box, each bounding box including a detected object of a predefined type; and

transmitting the plurality of crop regions;

wherein the method comprises either:

compiling a plurality of the crop regions into a composite view, and transmitting the composite view as a single frame of a single final video stream, wherein the relative spatial order of the objects of the one or more predefined types in the frame are maintained in a layout forming the frame of the final video stream, and wherein an arrangement of the layout is determined by an optimization process based on an aspect ratio of each crop region selected from the frame of the initial video stream and/or on the predefined preferred aspect ratio of each crop region; or

transmitting each of the crop regions as a separate final video stream, each final video stream including metadata containing information on spatial ordering, wherein each final video stream is advertised on a network as a virtual camera to which a subscription may be made.

2. The method of claim 1, wherein the step of detecting the objects of one or more predefined types in the frame is performed using a trained neural network, Haar Feature-based Cascade Classifiers, or a Histogram of Oriented Gradients.

3. The method according to any preceding claim, wherein the step of selecting a plurality of crop regions from the frame comprises:

reducing areas of the crop regions not overlapping any bounding boxes; and

reducing an overlap between the crop regions.

4. The method according to claim 3, wherein the step of selecting a plurality of crop regions from the frame of the initial video stream comprises parametrising each crop region by minimizing a cost function, wherein the cost function comprises a weighted sum of (i) a term correlated to areas of the crop regions not overlapping any bounding boxes; and (ii) a term correlated to the overlap between the crop regions.

5. The method according to claim 4, wherein the step of selecting a plurality of crop regions from the frame of the initial video stream comprises determining possible candidate sets of crop regions before minimizing the cost function, wherein the step of determining the possible candidate sets comprises one or more of:

identifying all possible candidate sets of crop regions that satisfy the predefined preferred aspect

ratio of each crop region;

limiting the possible candidate sets of crop regions to possible candidate sets where each crop region in the candidate set includes less than, or equal to, a predefined number of bounding boxes;

limiting the possible candidate sets of crop regions to possible candidate sets where all bounding boxes are included in the crop regions; and

limiting the possible candidate sets of crop regions such that the number of crop regions in each possible candidate set is:

if each crop region is transmitted as a separate frame of a separate final video stream ,less than, or equal to, a predefined maximum number of final video streams; or

if the crop regions are compiled into a composite view and the composite view is transmitted as a single frame of a single final video stream, less than, or equal to, a predefined maximum number of rectangles in a layout forming a composite view.

6. The method according to claim 4 or claim 5, wherein the cost function is optimized in a cost optimization process, with respect to one or more constraints relating to one or more of: a predefined maximum number of final video streams, a predefined maximum number of rectangles in a layout forming a composite view of crop regions, the number of bounding boxes, and a predefined preferred aspect ratio of each crop region.

7. The method according to any preceding claim, further comprising:

detecting objects of the one or more predefined types in a subsequent frame of the initial video stream;

setting a bounding box around the extent of each of the detected objects of the one or more predefined types in the subsequent frame of the initial video stream; and

selecting crop regions from the subsequent frame of the initial video, based on the crop regions from the frame, and movement of the bounding boxes between the frame and the subsequent frame.

8. The method according to claim 7, further comprising:

determining an amount of overlap between a pair of crop regions from the subsequent frame; and

merging the pair of crop regions from the subsequent frame, if the amount of overlap exceeds a predefined overlap threshold.

9. The method according to claim 7 or claim 8, further comprising:

determining a size of a crop region from the subsequent frame; and

splitting the crop region from the subsequent frame into two or more resulting crop regions if the size exceeds a predefined size threshold.

10. A data processing device for manipulating an initial video stream captured by a camera in a video conferencing endpoint into multiple views corresponding to regions of interest in accordance with the method of any preceding claim, the data processing device comprising:

a receiver configured to receive a frame of the initial video stream captured by the camera;

an object detection unit configured to detect objects of one or more predefined types in the frame of the initial video stream;

a cropping unit configured to select a plurality of crop regions from the frame of the initial video stream, each crop region including at least one bounding box, each bounding box including a detected object of a predefined type; and

a transmitter configured to transmit the plurality of crop regions either:

as a composite view in a single frame of a final video stream, the composite view formed by compiling the plurality of crop regions into the composite view, wherein the relative spatial order of the objects of the one or more predefined types in the frame are maintained in a layout forming the frame of the final video stream, and wherein an arrangement of the layout is determined by an optimization process based on an aspect ratio of each crop region selected from the frame of the initial video stream and/or on the predefined preferred aspect ratio of each crop region; or

as separate final video streams, each final video stream including metadata containing information on spatial ordering, wherein each final video stream is advertised on a network as a virtual camera to which a subscription may be made.

11. The data processing device of claim 10, wherein:

the receiver is configured to receive the initial video stream, wirelessly or via a wired connection, from the video conferencing endpoint having the camera or directly from the camera; and

the transmitter is configured to transmit, wirelessly or via a wired connection, the plurality of crop regions to one or more video conferencing endpoints for rendering and display at/by the one or more video conferencing endpoints.

12. A video conferencing system for manipulating an initial video stream into multiple views corresponding to regions of interest, the system comprising:

a camera configured to capture an initial video stream;

one or more video conferencing endpoints; and

a data processing device configured to perform the method of any of claims 1-9, the data processing device comprising:

a receiver configured to receive a frame of the initial video stream captured by the camera;

an object detection unit configured to detect objects of one or more predefined types in the frame of the initial video stream;

a cropping unit configured to select a plurality of crop regions from the frame of the initial video stream, each crop region including at least one bounding box, each bounding box including a detected object of a predefined type; and

a transmitter configured to transmit the plurality of crop regions to the one or more video conferencing endpoints either:

as a composite view in a single frame of a final video stream, the composite view formed by compiling the plurality of crop regions into the composite view, wherein the relative spatial order of the objects of the one or more predefined types in the frame are maintained in a layout forming the frame

of the final video stream, and wherein an arrangement of the layout is determined by an optimization process based on an aspect ratio of each crop region selected from the frame of the initial video stream and/or on the predefined preferred aspect ratio of each crop region; or

as separate final video streams, each final video stream including metadata containing information on spatial ordering, wherein each final video stream is advertised on a network as a virtual camera to which a subscription may be made.

13. The system of claim 12, wherein the one or more video conferencing endpoints are configured to render and display the crop regions as separate frames of separate final video streams, or in a composite view in a single frame of a final video stream.