

## CLAIMS

1. A video processing method of detecting visual relations in a video sequence, the method comprising:

decomposing the video sequence into a plurality of segments, each segment comprising a plurality of successive frames corresponding to a temporal portion of the video sequence and each segment overlapping with one or more neighboring segments by at least one frame;

for each segment,

detecting objects in frames of the segment;

tracking the detected objects over the segment to form a set of object tracklets for the segment;

for the detected objects, extracting object features;

for pairs of detected objects wherein each pair of detected objects comprises a first object which serves as a subject and a second object which serves as an object ~~tracklets of the set of object tracklets~~, extracting relativity features indicative of a relation between the subject and the objects ~~corresponding to the pair of object tracklets~~;

forming relation feature vectors for pairs of detected objects in the segment ~~tracklets~~ using the object features of objects corresponding to respective pairs of object tracklets and the relativity features of the respective pairs of detected objects ~~tracklets~~; and

generating a set of segment relation prediction results for the segment from the relation features vectors;

generating a set of visual relation instances, corresponding to pairs of subject and object, for the video sequence by merging the segment relation prediction results from different segments; and

generating a set of visual relation detection results, corresponding to pairs of subject and object, from the set of visual relation instances.

2. A method according to claim 1, wherein each segment has a length of frames L and the length of overlap is L/2 ~~the plurality of segments are overlapping segments~~.

3. A method according to claim 1 or claim 2, wherein the visual relation detection results comprise a set of indications of a relation subject, a relation predicate and a relation object, wherein the relation subject and the relation object are detected objects and the relation predicate indicates a relation between the relation subject and the relation object.
4. A method according to any preceding claim, wherein the object tracklets comprise a bounding box in each of a plurality of frames of the segment and a trajectory for the bounding box across the plurality of frames of the segment.
5. A method according to any preceding claim, wherein the relation between the subject and the objects corresponding to the pair of object tracklets, comprises at least one of the relative size, the relative position and the motion between the subject and object~~pair of objects corresponding to the pair of object tracklets~~.
6. A method according to any preceding claim, wherein generating the set of visual relation detection results from the set of visual relation instances comprises ranking the visual relation instances according to a confidence score and selecting the visual relation instances having the highest confidence scores as the visual relation detection results.
7. A method according to any preceding claim, wherein merging the segment relation prediction results from different segments comprises applying a greedy algorithm to merge pairs of object tracklets across successive segments.
8. A method according to any preceding claim, wherein generating the set of segment relation prediction results from the relation features vectors comprises applying a softmax function to set of relation feature vectors.
9. A method according to any preceding claim, wherein detecting objects in frames of the segment comprises detecting objects belonging to one of a set of pre-defined categories.

10. A method according to any preceding claim, wherein the visual relations comprise at least one dynamic relation which changes across the video sequence.

11. A computer readable medium carrying processor executable instructions which when executed on a processor cause the processor to carry out a method according to any one of claims 1 to 10.

12. A video processing system comprising a processor and a data storage device, the data storage device storing computer executable instructions operable by the processor to:

decompose a video sequence into a plurality of segments, each segment comprising a plurality of successive frames corresponding to a temporal portion of the video sequence and each segment overlapping with one or more neighboring by at least one frame;

for each segment,

detect objects in frames of the segment;

track the detected objects over the segment to form a set of object tracklets for the segment;

for the detected objects, extract object features;

for pairs of detected objects wherein each pair of detected objects comprises a first object which serves as a subject and a second object which serves as an object~~tracklets of the set of object tracklets~~, extract relativity features indicative of a relation between the subject and the objects~~corresponding to the pair of object tracklets~~;

form relation feature vectors for pairs of detected objects~~tracklets in the segment~~ using the object features of objects corresponding to respective pairs of object tracklets and the relativity features of the respective pairs of detected objects~~tracklets~~; and

generate a set of segment relation prediction results for the segment from the relation features vectors;

generate a set of visual relation instances, corresponding to pairs of subject and object, for the video sequence by merging the segment relation prediction results from different segments; and

generate a set of visual relation detection results, corresponding to pairs of subject and object, from the set of visual relation instances.

13. A video processing system according to claim 12, wherein each segment has a length of frames L and the length of overlap is L/2~~the plurality of segments are overlapping segments~~.

14. A video processing system according to claim 12 or 13, wherein the visual relation detection results comprise a set of indications of a relation subject, a relation predicate and a relation object, wherein the relation subject and the relation object are detected objects and the relation predicate indicates a relation between the relation subject and the relation object.

15. A video processing system according to any one of claims 12 to 14, wherein the object tracklets comprise a bounding box in each of a plurality of frames of the segment and a trajectory for the bounding box across the plurality of frames of the segment.

16. A video processing system according to any one of claims 12 to 15, wherein the relation between the subject and the objects corresponding to the pair of object tracklets, comprises at least one of the relative size, the relative position and the motion between the subject and ~~the pair of objects corresponding to the pair of object tracklets~~.

17. A video processing system according to any one of claims 12 to 16, wherein the data storage device comprises instructions operable by the processor to generate the set of visual relation detection results from the set of visual relation instances by ranking the visual relation instances according to a confidence score and selecting the visual relation instances having the highest confidence scores as the visual relation detection results.

18. A video processing system according to any one of claims 12 to 17, wherein the data storage device comprises instructions operable by the processor to merge

the segment relation prediction results from different segments by applying a greedy algorithm to merge pairs of object tracklets across successive segments.

19. A video processing system according to any one of claims 12 to 18, wherein the data storage device comprises instructions operable by the processor to generate the set of segment relation prediction results from the relation features vectors by applying a softmax function to set of relation feature vectors.

20. A video processing system according to any one of claims 12 to 18, wherein the data storage device comprises instructions operable by the processor to detect objects in frames of the segment by detecting objects belonging to one of a set of pre-defined categories.