State Of The Art Multi Object Tracking: An Overview

Vision and Cognitive Services

Gianmarco Cracco, Alessandro Manente
18th September 2020

UNIVERSITÀ DEGLI STUDI DI PADOVA

Overview

UNIVERSITÀ DEGLI STUDI DI PADOVA

- MOT has the goal to estimate trajectories of multiple moving objects. Here we present the SOTA model.
- Up until now $\longrightarrow$ best performance were obtained through Two-step (Object detection + Re-ID) methods high accuracy but reduced speed;
- One-step methods have risen: they share most feature across all the model $\rightarrow$ high speed but low accuracy.

Introduction

- MOT has the goal to estimate trajectories of multiple moving objects. Here we present the SOTA model.
- Up until now $\longrightarrow$ best performance were obtained through Two-step (Object detection + Re-ID) methods high accuracy but reduced speed;
- One-step methods have risen: they share most feature across all the model $\rightarrow$ high speed but low accuracy.

Introduction

- MOT has the goal to estimate trajectories of multiple moving objects. Here we present the SOTA model.
- Up until now $\longrightarrow$ best performance were obtained through Two-step (Object detection + Re-ID) methods high accuracy but reduced speed;
- One-step methods have risen: they share most feature across all the model $\rightarrow$ high speed but low accuracy.

One-Step method's Pt.1: problems

UNIVERSITÀ DEGLI STUDI DI PADOVA

One-Step methods presented prior to this model suffer from the following problems:

1. **Use of Anchors:** not suited for learning Re-ID features; multiple anchors responsible for the identity of the same object + feature map is downsampled by 8 times to balance accuracy and speed → too rough for Re-ID.
2. **Lack of Multi-Layer Feature Aggregation:** required to be able to handle both small and large object. → reduced number of identity switches.
3. **Dimensionality of the Re-ID Features:** usually high-dimensional features are used, but lower-dimensional features are better for MOT → reduced risk of overfitting.

One-Step method's Pt.1: problems

One-Step methods presented prior to this model suffer from the following problems:

1. **Use of Anchors:** not suited for learning Re-ID features; multiple anchors responsible for the identity of the same object + feature map is downsampled by 8 times to balance accuracy and speed $\rightarrow$ too rough for Re-ID.
2. **Lack of Multi-Layer Feature Aggregation:** required to be able to handle both small and large object. $\rightarrow$ reduced number of identity switches.
3. **Dimensionality of the Re-ID Features:** usually high-dimensional features are used, but lower-dimensional features are better for MOT $\rightarrow$ reduced risk of overfitting.

One-Step method's Pt.1: problems

Università degli Studi di Padova

One-Step methods presented prior to this model suffer from the following problems:

1. **Use of Anchors:** not suited for learning Re-ID features; multiple anchors responsible for the identity of the same object + feature map is downsampled by 8 times to balance accuracy and speed → too rough for Re-ID.
2. **Lack of Multi-Layer Feature Aggregation:** required to be able to handle both small and large object. → reduced number of identity switches.
3. **Dimensionality of the Re-ID Features:** usually high-dimensional features are used, but lower-dimensional features are better for MOT → reduced risk of overfitting.

One-Step method's Pt.1: problems

One-Step methods presented prior to this model suffer from the following problems:

1. **Use of Anchors:** not suited for learning Re-ID features; multiple anchors responsible for the identity of the same object + feature map is downsampled by 8 times to balance accuracy and speed → too rough for Re-ID.
2. **Lack of Multi-Layer Feature Aggregation:** required to be able to handle both small and large object. → reduced number of identity switches.
3. **Dimensionality of the Re-ID Features:** usually high-dimensional features are used, but lower-dimensional features are better for MOT → reduced risk of overfitting.

One-Step method's Pt.1: Solutions

Università degli Studi di Padova

To overcame all these problems → anchor-free object detection to estimate object centres on a high-resolution map, then parallel branch estimating pixel-wise Re-ID features to predict objects' identities.

The resulting model is NN based on ResNet-34 with the following changes/additions:

- Variant of Deep Layer Aggregation (DLA);
- Convolution layers are replaced by Deformable Convolution Layer;
- At the end are attached 4 heads responsible for the object detection and identity embedding.

One-Step method's Pt.1: Solutions

Università degli Studi di Padova

To overcame all these problems → anchor-free object detection to estimate object centres on a high-resolution map, then parallel branch estimating pixel-wise Re-ID features to predict objects' identities.

The resulting model is NN based on ResNet-34 with the following changes/additions:

- Variant of Deep Layer Aggregation (DLA);
- Convolution layers are replaced by Deformable Convolution Layer;
- At the end are attached 4 heads responsible for the object detection and identity embedding.

One-Step method's Pt.1: Solutions

To overcame all these problems → anchor-free object detection to estimate object centres on a high-resolution map, then parallel branch estimating pixel-wise Re-ID features to predict objects' identities.

The resulting model is NN based on ResNet-34 with the following changes/additions:

- Variant of Deep Layer Aggregation (DLA);
- Convolution layers are replaced by Deformable Convolution Layer;
- At the end are attached 4 heads responsible for the object detection and identity embedding.

One-Step method's Pt.1: Solutions

To overcame all these problems → anchor-free object detection to estimate object centres on a high-resolution map, then parallel branch estimating pixel-wise Re-ID features to predict objects' identities.

The resulting model is NN based on ResNet-34 with the following changes/additions:

- Variant of Deep Layer Aggregation (DLA);
- Convolution layers are replaced by Deformable Convolution Layer;
- At the end are attached 4 heads responsible for the object detection and identity embedding.

One-Step method's Pt.1: Solutions

To overcame all these problems → anchor-free object detection to estimate object centres on a high-resolution map, then parallel branch estimating pixel-wise Re-ID features to predict objects' identities.

The resulting model is NN based on ResNet-34 with the following changes/additions:

- Variant of Deep Layer Aggregation (DLA);
- Convolution layers are replaced by Deformable Convolution Layer;
- At the end are attached 4 heads responsible for the object detection and identity embedding.

Previous Works

- **Two-Step method:** object detection and Re-ID treated as 2 different tasks.
  CNN localizes objects, then an identity embedding network extracts Re-ID features and link the boxes $\rightarrow$ IoU, Kalman Filter and Hungarian Algorithm.
  PRO: best model for each task can be selected CON: slow performance.
- **One-Step Method:** share same weights and simultaneously accomplish object detection and Re-ID.
  PRO: reduced inference time CON: lower tracking accuracy than two-step due to anchor-based approach.

Previous Works

- **Two-Step method:** object detection and Re-ID treated as 2 different tasks.
  CNN localizes objects, then an identity embedding network extracts Re-ID features and link the boxes $\rightarrow$ IoU, Kalman Filter and Hungarian Algorithm.
  PRO: best model for each task can be selected CON: slow performance.
- **One-Step Method:** share same weights and simultaneously accomplish object detection and Re-ID.
  PRO: reduced inference time CON: lower tracking accuracy than two-step due to anchor-based approach.

Dataset Pt.1

The data used to obtain SOTA results on MOT challenge are:

- ETH, CityPerson, CalTech, MOT17, CUHK-SYSU and PRW.

Each dataset made of train and validation set as sequence of images + attached text file with the centre and dimensions of the bounding box for each object of interest contained.

Dataset Pt.1

The data used to obtain SOTA results on MOT challenge are:

- ETH, CityPerson, CalTech, MOT17, CUHK-SYSU and PRW.

Each dataset made of train and validation set as sequence of images + attached text file with the centre and dimensions of the bounding box for each object of interest contained.

Dataset pt.2

For our final experiments, we used 3 different videos:

- **Real-world scenario:** similar to original training data → limited duration (15 s) + lot of moving obstacles → suitable to test lack of long term memory.
- **Videogame:** medieval setting but still with a focus on people walking and standing in a city-like environment. Additional difficulty → costumes & ancient scenery.
- **2D animation:** objects are $2D$ characters walking on the road → environment similar to the training data.

Dataset pt.2

For our final experiments, we used 3 different videos:

- **Real-world scenario:** similar to original training data $\rightarrow$ limited duration (15 s) + lot of moving obstacles $\rightarrow$ suitable to test lack of long term memory.
- **Videogame:** medieval setting but still with a focus on people walking and standing in a city-like environment. Additional difficulty $\rightarrow$ costumes & ancient scenery.
- **2D animation:** objects are 2D characters walking on the road $\rightarrow$ environment similar to the training data.

Dataset pt.2

For our final experiments, we used 3 different videos:

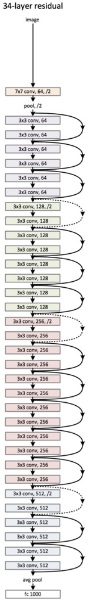
- **Real-world scenario:** similar to original training data $\rightarrow$ limited duration (15 s) + lot of moving obstacles $\rightarrow$ suitable to test lack of long term memory.
- **Videogame:** medieval setting but still with a focus on people walking and standing in a city-like environment. Additional difficulty $\rightarrow$ costumes & ancient scenery.
- **2D animation:** objects are $2D$ characters walking on the road $\rightarrow$ environment similar to the training data.

Dataset pt.2

For our final experiments, we used 3 different videos:

- **Real-world scenario:** similar to original training data $\rightarrow$ limited duration (15 s) + lot of moving obstacles $\rightarrow$ suitable to test lack of long term memory.
- **Videogame:** medieval setting but still with a focus on people walking and standing in a city-like environment. Additional difficulty $\rightarrow$ costumes & ancient scenery.
- **2D animation:** objects are $2D$ characters walking on the road $\rightarrow$ environment similar to the training data.

Backbone Pt.1



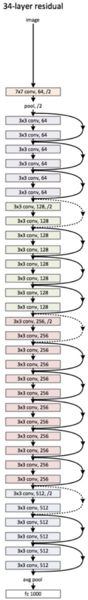
ResNet34 as basis $\rightarrow$ balance between speed-accuracy with combination of a deep architecture and skip-connections between layers.

Structure:

- (CONV7,BN,MAXPOOL3)
- (CONV3,BN,RELU)*3
- (CONV3,BN,RELU)*4
- (CONV3,BN,RELU)*6
- (CONV3,BN)*3
- AVGPOOL, FC, SOFTMAX

Backbone Pt.1

UNIVERSITÀ DEGLI STUDI DI PADOVA

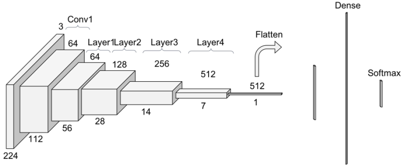


ResNet34 as basis → balance between speed-accuracy with combination of a deep architecture and skip-connections between layers.

Structure:

- (CONV7,BN,MAXPOOL3)
- (CONV3,BN,RELU)*3
- (CONV3,BN,RELU)*4
- (CONV3,BN,RELU)*6
- (CONV3,BN)*3
- AVGPOOL, FC, SOFTMAX

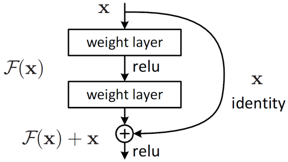
Backbone Pt.2



From one block to the next one, to downsample → NO POOL but Stride=2.

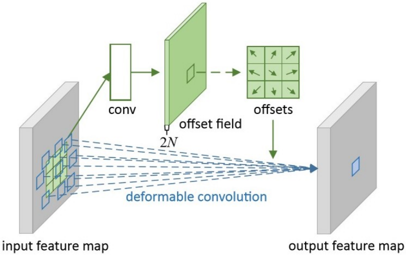
Number of filter per each block doubled to preserve time complexity.

Every 2 CONV → Identity mapping.



Deformable Convoltional Layers

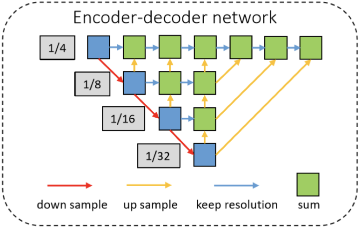
First alteration $\rightarrow$ convolutional layers substituted with deformable convolutional layers to increase the capabilities of the model to adapt to variations in scales and poses of the objects.



Offsets obtained with CONV over the same input feature map.

Conv kernels & offsets are learned simultaneously.

Second alteration $\rightarrow$ variant of Deep Layer Aggregation (DLA) has been applied.



Both IDA and HDA are used + additional skip connection w.r.t. original implementation to accommodate objects of different scales.

Object Detection Branch Heads

Università degli Studi di Padova

The 3 Object Detection heads are the following:

1. **Heatmap Head:** estimates the locations of object centres using heatmap based representation. The response in a given location is expected to be 1 if it coincides with GT and decays exponentially with the increase of the distance between the object centre and the heatmap location.
2. **Centre Offset Head:** improves accuracy on the estimation of localization of the objects → critical to achieve and sustain good performance of Re-ID.
3. **Box Size Head:** estimates the dimensions of the object bounding box. Not directly related to Re-ID operations.

Object Detection Branch Heads

Università degli Studi di Padova

The 3 Object Detection heads are the following:

1. **Heatmap Head:** estimates the locations of object centres using heatmap based representation. The response in a given location is expected to be 1 if it coincides with GT and decays exponentially with the increase of the distance between the object centre and the heatmap location.
2. **Centre Offset Head:** improves accuracy on the estimation of localization of the objects → critical to achieve and sustain good performance of Re-ID.
3. **Box Size Head:** estimates the dimensions of the object bounding box. Not directly related to Re-ID operations.

Object Detection Branch Heads

The 3 Object Detection heads are the following:

1. **Heatmap Head:** estimates the locations of object centres using heatmap based representation. The response in a given location is expected to be 1 if it coincides with GT and decays exponentially with the increase of the distance between the object centre and the heatmap location.
2. **Centre Offset Head:** improves accuracy on the estimation of localization of the objects $\rightarrow$ critical to achieve and sustain good performance of Re-ID.
3. **Box Size Head:** estimates the dimensions of the object bounding box. Not directly related to Re-ID operations.

Object Detection Branch Heads

The 3 Object Detection heads are the following:

1. **Heatmap Head:** estimates the locations of object centres using heatmap based representation. The response in a given location is expected to be 1 if it coincides with GT and decays exponentially with the increase of the distance between the object centre and the heatmap location.
2. **Centre Offset Head:** improves accuracy on the estimation of localization of the objects $\rightarrow$ critical to achieve and sustain good performance of Re-ID.
3. **Box Size Head:** estimates the dimensions of the object bounding box. Not directly related to Re-ID operations.

Identity Embedding Branch Head

Identity Embedding Branch: only one head, implemented with a convolutional layer with 128 kernels to extract identity embedding features for each location.

$$\Downarrow$$

Responsible for generating features that can distinguish different objects.

Identity Embedding Branch Head

Identity Embedding Branch: only one head, implemented with a convolutional layer with 128 kernels to extract identity embedding features for each location.

$$\Downarrow$$

Responsible for generating features that can distinguish different objects.

The Model

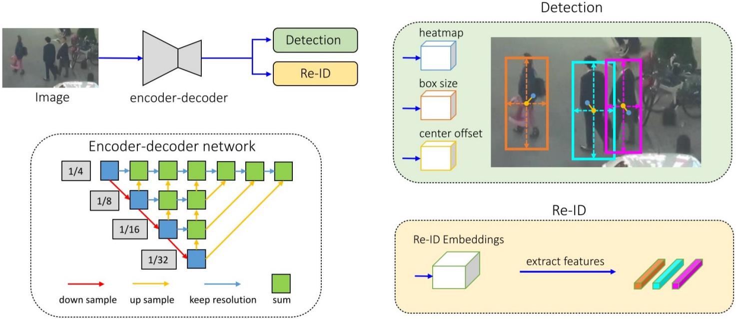


Figure: Input is given to an encoder-decoder network to extract high resolution feature map, with a stride of 4. This outputs to two parallel heads to predict the bounding boxes and the Re-ID features. The predicted objects are fed to standard box linking techniques.

- For each box $b^i = \left(x_1^i, y_1^i, x_2^i, y_2^i\right)$ in the image, the object centre $\left(c_x^i, c_y^i\right) = \left(\frac{x_1^i + x_2^i}{2}, \frac{y_1^i + y_2^i}{2}\right)$ is computed and the location on the heatmap $\left(\widetilde{c}_x^i, \widetilde{c}_y^i\right)$ is obtained dividing the stride of 4.
- The heatmap response at location (x, y) is $M_{xy} = \sum_{i=1}^{N} \exp^{-\frac{\left(x - i_x^i\right)^2 + \left(y - \tilde{c}_y^i\right)^2}{2\sigma_c^2}}$ with N number of objects in the image and σ_c the standard deviation.

Heatmap Loss pt.1

- For each box $b^i = (x_1^i, y_1^i, x_2^i, y_2^i)$ in the image, the object centre $\left(c_x^i, c_y^i\right) = \left(\frac{x_1^i + x_2^i}{2}, \frac{y_1^i + y_2^i}{2}\right)$ is computed and the location on the heatmap $\left(\widetilde{c}_x^i, \widetilde{c}_y^i\right)$ is obtained dividing the stride of 4.
- The heatmap response at location (x, y) is $M_{xy} = \sum_{i=1}^{N} \exp^{-\frac{\left(x - i_x^i\right)^2 + \left(y - \tilde{c}_y^i\right)^2}{2\sigma_c^2}}$ with N number of objects in the image and σ_c the standard deviation.

Heatmap Loss pt.2

Resulting loss → pixel-wise logistic regression with focal loss:

$$L_{\text{heatmap}} = -\frac{1}{N}\sum_{xy} \begin{cases} \left(1-\hat{M}_{xy}\right)^{\alpha} \log\left(\hat{M}_{xy}\right) & \text{if } M_{xy}=1 \\ \left(1-M_{xy}\right)^{\beta}\left(\hat{M}_{xy}\right)^{\alpha}\log\left(1-\hat{M}_{xy}\right) & \text{otherwise} \end{cases}$$

$\hat{M}$ is the estimated heatmap while α, β are the parameters.
The focal loss is:

$$FL\left(p\right) = -\left(1-p\right)^{\gamma}\log\left(p\right)$$

Offset and Size Loss

- For each box $b^i = (x_1^i, y_1^i, x_2^i, y_2^i)$ the size is

$$s^i = \left(x_2^i - x_1^i, y_2^i - y_1^i\right),$$

the offset is

$$o^i = \left(\frac{c_x^i}{4}, \frac{c_y^i}{4}\right) - \left(\left\lfloor \frac{c_x^i}{4} \right\rfloor, \left\lfloor \frac{c_y^i}{4} \right\rfloor\right).$$

- Using l_1 norm, the final loss for the two heads is:

$$L_{\text{box}} = \sum_{i=1}^{N} \left(\left\|o^i - \hat{o}^i\right\|_1 + \left\|s^i - \hat{s}^i\right\|_1\right)$$

Offset and Size Loss

- For each box $b^i = (x_1^i, y_1^i, x_2^i, y_2^i)$ the size is

$$s^i = \left(x_2^i - x_1^i, y_2^i - y_1^i\right),$$

the offset is

$$o^i = \left(\frac{c_x^i}{4}, \frac{c_y^i}{4}\right) - \left(\left\lfloor \frac{c_x^i}{4} \right\rfloor, \left\lfloor \frac{c_y^i}{4} \right\rfloor\right).$$

- Using l_1 norm, the final loss for the two heads is:

$$L_{\text{box}} = \sum_{i=1}^{N} \left(\left\| o^i - \hat{o}^i \right\|_1 + \left\| s^i - \hat{s}^i \right\|_1 \right)$$

Identity Embedding Loss

UNIVERSITÀ DEGLI STUDI DI PADOVA

This task is treated as a classification one.

- All objects instances with same identity belongs to one class. For each box b^i in the image a object centre on the heatmap is obtained. A feature vector E_{x^i,y^i} is extracted at location and is mapped to a class distribution vector $p(k)$.
- Given $L^i(k)$ the one-hot representation of GT class label, the softmax loss is computed:

$$L_{\text{identity}} = -\sum_{i=1}^{N}\sum_{k=1}^{K} L^i(k)\log(p(k))$$

with K number of classes.

Identity Embedding Loss

This task is treated as a classification one.

- All objects instances with same identity belongs to one class. For each box b^i in the image a object centre on the heatmap is obtained. A feature vector E_{x^i,y^i} is extracted at location and is mapped to a class distribution vector $p(k)$.
- Given $L^i(k)$ the one-hot representation of GT class label, the softmax loss is computed:

$$L_{\text{identity}} = -\sum_{i=1}^{N}\sum_{k=1}^{K} L^i(k)\log(p(k))$$

with K number of classes.

Online Linking Pt.1: Inference

UNIVERSITÀ DEGLI STUDI DI PADOVA

Inference: the input is an image of size 1088x608.

$\Downarrow$

Given the predicted heatmap, with Non-Maximum Suppression peak keypoints are extracted; only those with scores larger than a threshold are kept.

$\Downarrow$

Bounding boxes are computed and Identity Embeddings extracted.

Online Linking Pt.1: Inference

Inference: the input is an image of size 1088x608.

$\Downarrow$

Given the predicted heatmap, with Non-Maximum Suppression peak keypoints are extracted; only those with scores larger than a threshold are kept.

$\Downarrow$

Bounding boxes are computed and Identity Embeddings extracted.

Online Linking Pt.1: Inference

Inference: the input is an image of size 1088x608.

$$\Downarrow$$

Given the predicted heatmap, with Non-Maximum Suppression peak keypoints are extracted; only those with scores larger than a threshold are kept.

$$\Downarrow$$

Bounding boxes are computed and Identity Embeddings extracted.

Online Linking Pt.2: Box Linking

UNIVERSITÀ DEGLI STUDI DI PADOVA

Box Linking: tracklets are initialized based on the estimated boxes

$\Downarrow$

Boxes are linked to existing tracklets according to distances measured by Re-ID features and IoU's.

$\Downarrow$

Kalman filter is used to predict locations of tracklets in current frame. If the distance too big, the cost is set to $\infty \rightarrow$ prevents from linking the detections with large motion.
To handle appearance variations such features are updated at each time step.

Online Linking Pt.2: Box Linking

Università degli Studi di Padova

Box Linking: tracklets are initialized based on the estimated boxes

$\Downarrow$

Boxes are linked to existing tracklets according to distances measured by Re-ID features and IoU's.

$\Downarrow$

Kalman filter is used to predict locations of tracklets in current frame. If the distance too big, the cost is set to $\infty \rightarrow$ prevents from linking the detections with large motion.
To handle appearance variations such features are updated at each time step.

Online Linking Pt.2: Box Linking

Box Linking: tracklets are initialized based on the estimated boxes

$$\Downarrow$$

Boxes are linked to existing tracklets according to distances measured by Re-ID features and IoU's.

$$\Downarrow$$

Kalman filter is used to predict locations of tracklets in current frame. If the distance too big, the cost is set to $\infty \rightarrow$ prevents from linking the detections with large motion.
To handle appearance variations such features are updated at each time step.

Failures

Our primary efforts → fine-tune the model in order to obtain new and different results: play with both the parameters and hyper-parameters of heads' final layers to reduce training time.

Layers were frozen but continuous GPU error → impossible to train our model.
In addition, to get test performance, results set must be uploaded to MOTchallenge.net with following restrictions:

- 3 days between an upload and the next one;
- Max 4 attempts.

⇓

We had to give up on trying.

Failures

Our primary efforts → fine-tune the model in order to obtain new and different results: play with both the parameters and hyper-parameters of heads' final layers to reduce training time.

Layers were frozen but continuous GPU error → impossible to train our model.
In addition, to get test performance, results set must be uploaded to MOTchallenge.net with following restrictions:

- 3 days between an upload and the next one;
- Max 4 attempts.

⇓

We had to give up on trying.

Failures

Our primary efforts → fine-tune the model in order to obtain new and different results: play with both the parameters and hyper-parameters of heads' final layers to reduce training time.

Layers were frozen but continuous GPU error → impossible to train our model.
In addition, to get test performance, results set must be uploaded to MOTchallenge.net with following restrictions:

- 3 days between an upload and the next one;
- Max 4 attempts.

⇓

We had to give up on trying.

With `demo` function provided by the code → we fed the model with 3 videos previously announced:

- **Real-world scenario footage:** shows main problem with the model → lack of long-term memory, whenever something covers the woman standing in the background her ID changes.

  Changing the number of frames after which the linking algorithm delete a given unused tracklet → reduce the amount of ID switch changes for the woman in the background to 3.

  No appreciable results were obtained choosing a buffer size greater than 90 frames, in fact it would always switch the ID two times → need for better technique for linking with a focus on the memory of it.

With `demo` function provided by the code → we fed the model with 3 videos previously announced:

- **Real-world scenario footage:** shows main problem with the model → lack of long-term memory, whenever something covers the woman standing in the background her ID changes.

  Changing the number of frames after which the linking algorithm delete a given unused tracklet → reduce the amount of ID switch changes for the woman in the background to 3.

  No appreciable results were obtained choosing a buffer size greater than 90 frames, in fact it would always switch the ID two times → need for better technique for linking with a focus on the memory of it.

Experiments

With `demo` function provided by the code → we fed the model with 3 videos previously announced:

- **Real-world scenario footage:** shows main problem with the model → lack of long-term memory, whenever something covers the woman standing in the background her ID changes.

  Changing the number of frames after which the linking algorithm delete a given unused tracklet → reduce the amount of ID switch changes for the woman in the background to 3.

  No appreciable results were obtained choosing a buffer size greater than 90 frames, in fact it would always switch the ID two times → need for better technique for linking with a focus on the memory of it.

Woman in Background

Figure: Evidences of lack of long term memory of the model. The woman standing in the background is here shown after her figure was covered by different object obstructing the view of the camera and it can be seen her ID switching in the sequence 10-12-14-24.

Videogame

Videogame footage: lack of contrast and lack of difference in colours between people and background → people often not recognised or switched IDs. Works quite well given the fact that the model never saw a clip coming from a videogame.

Animation

2D animation footage: the model can follow $2D$ animated characters quite well, and shows problems only when the pose of characters is too far from the standing one.

Lack of memory: future proposal Pt.1

Our proposal to fix the lack of long-term memory that affects the online linking algorithm:

- instead of relying on standard algorithms for tracking (IoU + Kalman filter + Hungarian algorithm) we propose the use of a NN for classification that takes as input the tracklets of the estimated boxes and classify them sequentially.
- If a tracklet has been classified as a never seen before class (no class reach a certain threshold of probability) $\rightarrow$ final layer is extended by 1 unit and random weights are initialised to that unit.
- To update the classification capabilities of the network to the new info, the final layer is fine-tuned using the tracklets of the last frame as training set.

Lack of memory: future proposal Pt.1

Our proposal to fix the lack of long-term memory that affects the online linking algorithm:

- instead of relying on standard algorithms for tracking (IoU + Kalman filter + Hungarian algorithm) we propose the use of a NN for classification that takes as input the tracklets of the estimated boxes and classify them sequentially.
- If a tracklet has been classified as a never seen before class (no class reach a certain threshold of probability) $\rightarrow$ final layer is extended by 1 unit and random weights are initialised to that unit.
- To update the classification capabilities of the network to the new info, the final layer is fine-tuned using the tracklets of the last frame as training set.

Lack of memory: future proposal Pt.1

Università degli Studi di Padova

Our proposal to fix the lack of long-term memory that affects the online linking algorithm:

- instead of relying on standard algorithms for tracking (IoU + Kalman filter + Hungarian algorithm) we propose the use of a NN for classification that takes as input the tracklets of the estimated boxes and classify them sequentially.
- If a tracklet has been classified as a never seen before class (no class reach a certain threshold of probability) $\rightarrow$ final layer is extended by 1 unit and random weights are initialised to that unit.
- To update the classification capabilities of the network to the new info, the final layer is fine-tuned using the tracklets of the last frame as training set.

Lack of memory: future proposal Pt.1

Our proposal to fix the lack of long-term memory that affects the online linking algorithm:

- instead of relying on standard algorithms for tracking (IoU + Kalman filter + Hungarian algorithm) we propose the use of a NN for classification that takes as input the tracklets of the estimated boxes and classify them sequentially.
- If a tracklet has been classified as a never seen before class (no class reach a certain threshold of probability) $\rightarrow$ final layer is extended by 1 unit and random weights are initialised to that unit.
- To update the classification capabilities of the network to the new info, the final layer is fine-tuned using the tracklets of the last frame as training set.

Lack of memory: future proposal Pt.2

- During the training through the loss, penalise the weights associated to classes that haven't been seen for at least a given number of frames;
- This number should be at least 60 (which equals to 2 s, assuming default frame rate is 30fps) if the aim of the model is still to track people in urban environment, due to possible occlusions;
- Furthermore to let the model forget about classes not present in the frames for a given threshold of time $\rightarrow$ weights corresponding to those old classes should be set equal to 0.

Lack of memory: future proposal Pt.2

- During the training through the loss, penalise the weights associated to classes that haven't been seen for at least a given number of frames;
- This number should be at least 60 (which equals to 2 s, assuming default frame rate is 30fps) if the aim of the model is still to track people in urban environment, due to possible occlusions;
- Furthermore to let the model forget about classes not present in the frames for a given threshold of time $\rightarrow$ weights corresponding to those old classes should be set equal to 0.

Lack of memory: future proposal Pt.2

- During the training through the loss, penalise the weights associated to classes that haven't been seen for at least a given number of frames;
- This number should be at least 60 (which equals to 2 s, assuming default frame rate is 30fps) if the aim of the model is still to track people in urban environment, due to possible occlusions;
- Furthermore to let the model forget about classes not present in the frames for a given threshold of time $\rightarrow$ weights corresponding to those old classes should be set equal to 0.

Conclusion

- We presented an overview of a one-shot anchor-free multiple object tracking.
  Reasons behind the failures of previous works $\longrightarrow$ anchor-based approach + Two-Shot methods.
- We showed the strength and also weakness of our model $\longrightarrow$ lack of long term memory. Our proposal: NN to classify each tracklet

Conclusion

Università degli Studi di Padova

- We presented an overview of a one-shot anchor-free multiple object tracking.
  Reasons behind the failures of previous works $\longrightarrow$ anchor-based approach + Two-Shot methods.
- We showed the strength and also weakness of our model $\longrightarrow$ lack of long term memory. Our proposal: NN to classify each tracklet

The End

Thank you for your attention!

IoU

prova

prova

prova

prova