A Deeper Dive into Multi-Face Tracking in Unconstrained Videos

Ariel Rakovitsky and Charles An

Background

Multi-object tracking is an exciting field of computer vision that involves not only object tracking as discussed in class but also data association techniques. The problem is simple: given a video of many subjects moving about, track them throughout the video. Data association is necessary to recognize a single subject between frames. A specific problem within this field is the challenge of camera angle switches mid-video. With this project we explore the field of multi-object tracking as it relates to face tracking in a video with many camera angle switches.

Approach

The current state-of-the-art in this field is the paper *A Priorless Method for Multi-Object Tracking* (https://ieeexplore.ieee.org/document/8578161). In short, the paper describes a method for tracking faces in a crowd in a video that involves camera angle switches. Unlike previous papers in the field, Lin and Hung's algorithm does not require any prior data about the video, like the number of subjects. The primary focus of our project was the reimplementation the algorithm described in the paper in an attempt to replicate the results of the authors. To do so, we implemented our own DPM model based on an open source pose estimator and used the VGG16 network as described in the paper.

The second part of our project involved taking a deeper dive into the core of a multi-object tracking system. The authors of the *Priorless* paper propose their own algorithm for tracklet linking and data association but rely on two external tools to do so. The first is a face detector (based on a DPM) used for identifying where a face is in the video. The second is a pretrained neural network for face feature extraction. We varied both the face detector (using a HOG face detector and a CNN-based face detector) and the neural network (using SENet50 and ResNet50). All three neural networks were pre-trained on the same data.

The dataset used for this task was the annotated music video dataset put forward in *Tracking Persons of Interest via Adaptive Discriminative Features* (https://link.springer.com/chapter/10.1007/978-3-319-46454-1_26).

Results

Table 1a: Average CLEAR MOT Metrics. Varying Face Detector

Model	Recall ↑	Precision ↑	F1 ↑	FAF ↓	MOTA↑	MOTP↑
CNN & ResNet	42.3	49.4	13.5	0.433	9.3	50.5
HOG & ResNet	41.3	46.0	16.3	0.472	10.5	**52.5**
DPM & ResNet	**78.1**	**87.4**	**17.2**	**0.283**	**62.4**	42.1
Lin & Hung	81.7	90.2	85.3	0.27	69.2	86.0

Table 1b: Average CLEAR MOT Metrics. Varying Feature Extractor

Model	Recall ↑	Precision ↑	F1 ↑	FAF ↓	MOTA↑	MOTP↑
DPM & ResNet	**78.1**	87.4	17.2	**0.283**	62.4	42.1
DPM & VGG	76.6	**88.9**	15.9	0.311	**72.3**	**47.1**
DPM & SENet	73.9	82.4	**19.2**	0.404	54.0	42.1
Lin & Hung	81.7	90.2	85.3	0.27	69.2	86.0

Table 2a: WCP Scores. Varying Face Detector

Model	Westlife	Pussycat Dolls	Hello Bubble	Girls Aloud
CNN & ResNet	0.97	0.96	0.69	0.87
HOG & ResNet	0.98	0.96	0.82	0.89
DPM & ResNet	0.895	0.868	0.982	0.904
Lin & Hung	0.86	0.79	0.70	0.92

Table 2b: WCP Scores. Varying Feature Extractor

Model	Westlife	Pussycat Dolls	Hello Bubble	Girls Aloud
DPM & ResNet	0.895	0.868	**0.982**	**0.904**
DPM & VGG	**0.933**	**0.876**	0.794	0.890
DPM & SENet	0.875	0.847	0.807	0.829
Lin & Hung	0.86	0.79	0.70	0.92

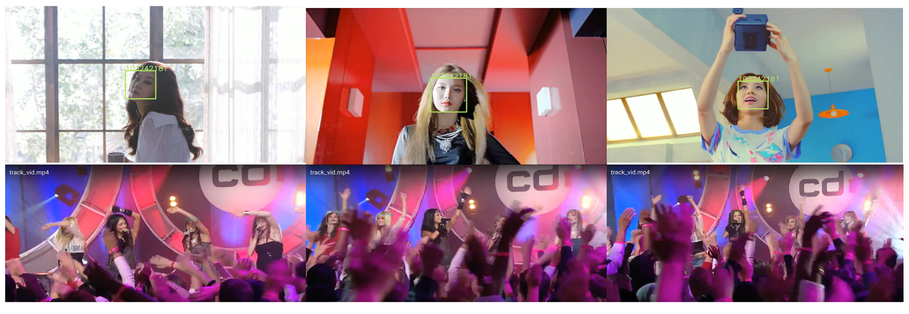
Table 3: DPM CLEAR MOT Metrics

Video	Recall ↑	Precision ↑	F1 ↑	FAF ↓	MOTA↑	MOTP↑
Apink	60.8	94.8	17.6	0.0589	55.3	33.0
Bruno Mars	72.7	87.9	12.8	0.362	58.9	43.0
Darling	92.7	91.2	11.3	0.2144	79.3	43.4
Girls Aloud	87.9	96.3	13.5	0.3224	57.0	46.4
Hello Bubble	71.2	96.9	16.6	0.0382	66.2	37.3
Pussycat Dolls	71.0	85.4	21.1	0.3224	57.0	46.4
Westlife	81.5	70.1	19.1	0.7302	45.6	43.7

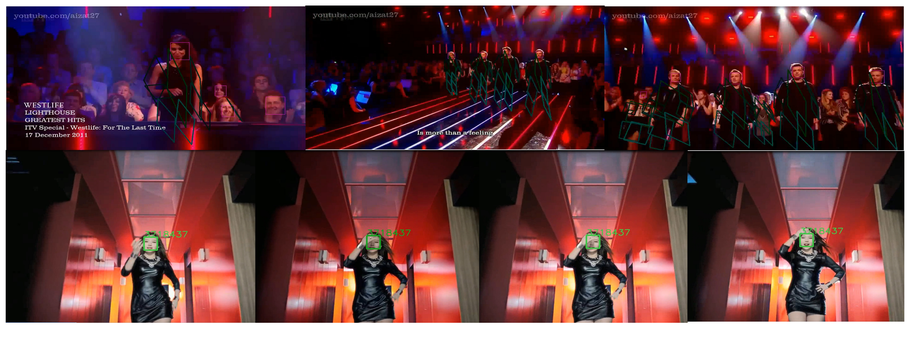
Tracklet Example

Reidentification and Linking

Reidentification and Linking



Difficulties and Challenges



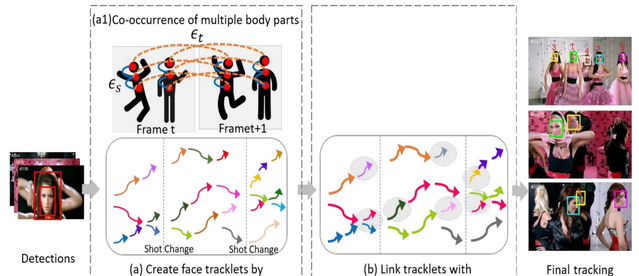
DPM Examples

Algorithm

1. Create tracklets: tracklets are short (about 10-100 frames) sequences of faces belonging to one person in the video. They are generated without feature extraction or linking. We used several approaches to generate tracklets:

1. DPM Model (paper implementation): Between two consecutive frames, if the overlap between corresponding body part bounding boxes is above a threshold, the two observations belong to the same tracklet.
2. Openpose Keypoints Model: Between two consecutive frames, if the overlap between corresponding body part keypoints is above a threshold, the observations belong to the same tracklet.
3. Face Only Model: Between two consecutive frames, if the overlap between two face bounding boxes is above a threshold, the observations belong to the same tracklet. Face detectors used are dlib-cnn and HOG.

2. Link tracklets: tracklets are linked on the basis of features associated faces in each tracklet. Features are generated using VGG16, ResNet50, or SENet 50. We link tracklets with strong associations and use the linked tracklets to construct a constrained graph for extracting clusters. Clusters in the graph form completed Tracks.



Evaluation

We used two metrics to quantitatively evaluate the performance of our model: one evaluating our clustering and another evaluating our tracking. To evaluate clustering, we used a Weighted Clustered Purity (WCP) metric to measure how well our model clustered faces based on identities. Tracks/Clusters containing faces from only or mostly the same person yield high WCP values.

To evaluate tracking, we used some of the most widely accepted evaluation metrics, the CLEAR MOT standards. The specific metrics used were Recall, Precision, F1, FAF, MOTA, and MOTP. The MOTA metric is calculed based on misses, false positives and mismatches.

Conclusion

We conclude that the deformable parts model has strong advantages over using convolutional neural network and histogram of oriented gradients face detectors in building tracklets based on an overlap algorithm. The deformable parts model creates longer tracklets and this proves to be very influential in the results of the model because the longer the tracklets are, the greater the number of possible facial variations of each person can be captured, thus improving the linking.

Changing the feature extractors displayed less variance than changing the face detectors, demonstrating that the most important part of the model is the tracklet generation.

Using a DPM, like Lin & Hung used, we were able to approximately replicate the results of the paper. Results had a strong dependence on the properties of the video that the system was run on.

Acknowledgements

We would like to thank Professor Olga Russakovsky and our project advisor Felix Yu.