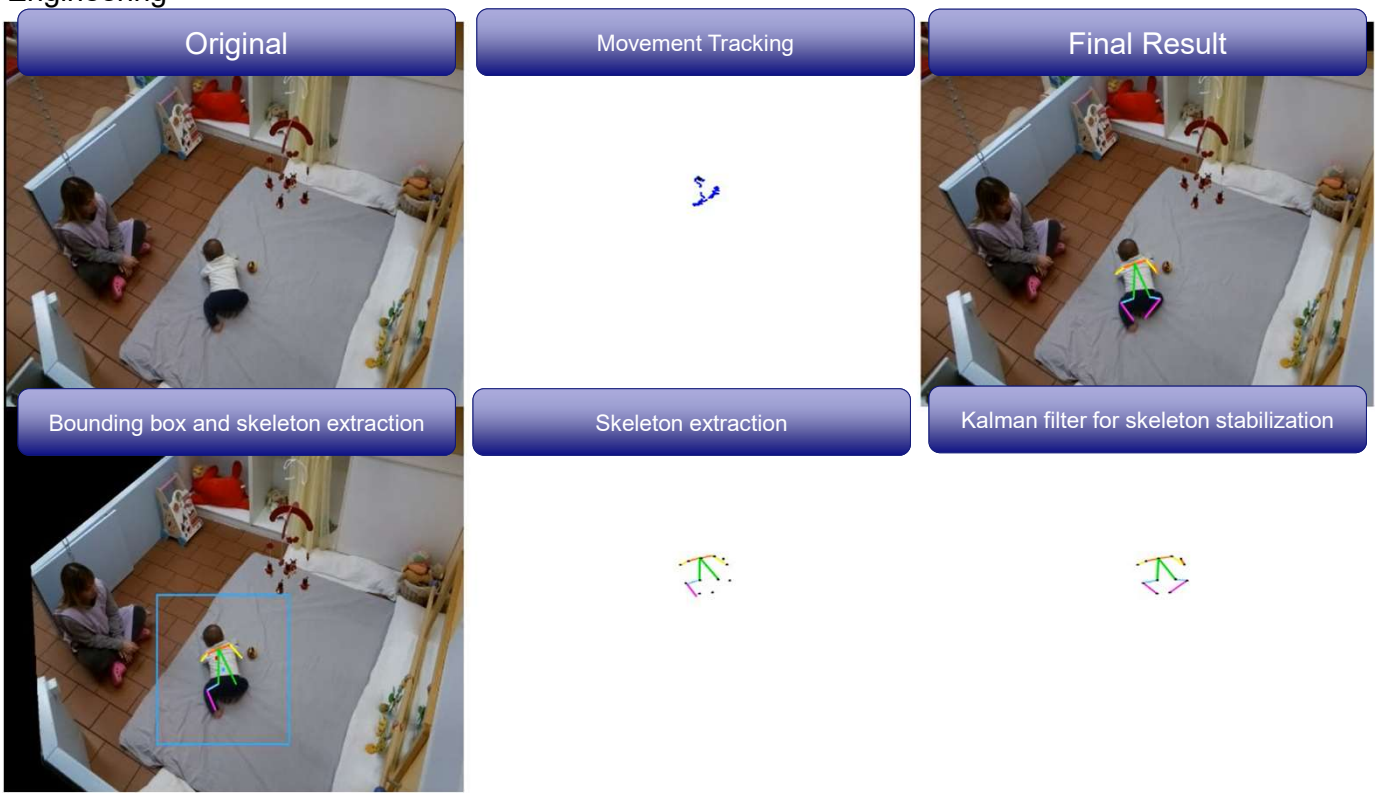
Scuola universitaria professionale della Svizzera italiana
Dipartimento tecnologie innovative

SUPSI

Infants Activity Recognition based on human pose estimation as a support for privacy-preserving neurodevelopmental disorders diagnosis

Studente	Relatore	Correlatore	Committente
Simone Sguazza	Michela Papandrea	-	-

Corso di laurea	Modulo	Anno	Data
Master of Science in Engineering	MP27_0001 - Thesis	2020	07 September 2020



Abstract

The project focuses on a Computer Vision application, which aims to help automation in the analysis of infant behaviors in order to help in the early diagnosis of neurological developmental disorders. The analyzed dataset corresponds to a set of videos of infants aged between 15 and 18 months, free to move inside a fixed multi-camera indoor environment, in an individual free play area with close interaction with the educator and other children outside the free play area. The context of this application is to help the doctors examine how infants use some specific toys to detect neurological developmental disorders. This task is mostly performed manually by the experts who monitor each video and make annotations, but is a time-consuming task. This application is designed to help experts automate this task. Infants are not aware of the cameras; they explore and interact with the environment and the adult. The video data is not collected for a Computer Vision application, the child can move freely and do whatever he or she wants. Unlike most State of Art studies, where the environment and the subject are in function of skeletal extraction to obtain a qualitative sample, nothing was done here to limit the external noise during the environmental preparation.

The project was divided into three main phases:

- Track a generic infant in a delimited free play area, where it plays with specific toys and/or explore the environment. I built a tailored Machine Learning Tracker to address the problem of aliasing introduced by the adult on the ground within the free play area.
- Generate the skeleton with OpenPose technology and stabilize it with a Linear Kalman filter.
- Extract the skeleton, encode the information of the three cameras within an image and infer the actions with a tailored CNN architecture classifier.

Goals

- Create a generic tracker
  - with only the knowledge of a generic child area distribution and the free-play area zone.
  - only one child and one adult can be within the ROI. The adult can go out and enter the play area during the video experience, S/he can interact with the child in any way he wants. Anything can happen, there are no constraints and the child is not aware of the camera.
  - in presence of deterministic aliasing problems due to a mirror and other human like puppets
  - in presence of stochastic aliasing problems due to adult on the ground and its interaction with the environment and with the child (every video is different)
  - robust to perspective of the camera
- Extract skeletons
  - avoiding false positive skeletons generated by OpenPose
  - choosing the best option among different skeleton proposals for each frame
  - stabilize the skeleton and infer missing parts
- Encode the information of the three cameras and infer the action of the infants.
  - Study the limits of the problem of recognition of human activities in a free play area, in a noisy environment, where the skeleton could be complete, partial or missing, with spurious limbs.

Conclusion

The generic tracker works quite well, but it is not always robust enough to always follow the child like a human. Some specific videos may have some track of the child delayed in some specific portion of the video. It is robust to occlusions, if it has lost the baby it will catch it after a few seconds. For some videos, the adult on ground alias is too strong in some specific part of the video and the tracker may lose the child for a few seconds. To avoid this, specific masks in the video are set to limit the influence of the specific alias portion by filtering them out.

Even with a perfect track, OpenPose may have problems estimating the specific position of the child's skeleton due to the camera angle and perspective. This can lead to incomplete, partial and/or spurious skeletons; in the worst cases the skeleton can be totally wrong, missing or missing the specific part of the body that performs the action.

The video ground truth of the actions is associated to the actions only with some specific toys, while all actions outside of this context are not mapped. Actions with these toys can be made according to the specific toys, the environment and/or the influence of external agents (adults and other children). Even selecting the good skeleton regions of each video and analyzing only the action well represented in the dataset, this huge variance of skeletons for the same action leads to a poor performance of the classifier.