Imperial College London

REAL-TIME MULTI-PERSON POSE TRACKING USING DATA ASSIMILATION

Caterina Buizza Tobias Fischer Yiannis Demiris

Personal Robotics Laboratory, Department of Electrical and Electronic Engineering

Imperial College London

WACV 2020

1 Introduction

We propose a framework for the integration of data assimilation and machine learning methods in human pose estimation, with the aim of enabling any pose estimation method to be run in real-time, whilst also increasing consistency and accuracy.

Our framework presents a real-time tracking module for any single or multi-person pose estimation system. Specifically, tracking is performed by a number of Kalman filters initiated for each new person appearing in a motion sequence. This permits tracking of multiple skeletons and reduces the frequency that computationally expensive pose estimation has to be run, enabling online pose tracking.

The main contributions of this paper are the following:

- We provide a practical example of how Data Assimilation and Machine Learning methods can be combined and used in a complementary manner.
- We implement a Kalman filter-based tracking module that can be applied to any human motion estimation algorithm, decreasing the average run time per frame and increasing the consistency of joint position estimation.
- We perform a noise analysis on three motion estimation systems to obtain better measurement and noise covariance matrices for the filtering module.
- We implement an identification algorithm to improve consistency of labelling of skeletons between frames.

2 Method: Kalman Filters

Kalman filters are powerful tools for dealing with measurement uncertainties and data inconsistencies. The equations of a Kalman filter also allow the definition of a dynamical model to describe the state evolution between timesteps.

The general time update equations for a discrete Kalman filter are

$$\hat{x}^-_{k+1} = \mathbf{A}_k \hat{x}_k + \mathbf{B}_k u_k + w_k \quad (1)$$

$$\mathbf{P}^-_{k+1} = \mathbf{A}_k \mathbf{P}_k \mathbf{A}^T_k + \mathbf{Q}_k \quad (2)$$

where $\mathbf{A}$ is the state transition matrix, $\mathbf{B}$ is the input matrix, w is the process noise, and $\mathbf{P}, \mathbf{Q}$ are the estimate and process covariance matrices respectively. The values n and l correspond to the state and input dimensions respectively. To these equations, we add the filter measurement update equations:

$$\mathbf{K}_k = \mathbf{P}^-_k \mathbf{H}^T_k (\mathbf{H}_k \mathbf{P}^-_k \mathbf{H}^T_k + \mathbf{R}_k)^{-1} \quad (3)$$

$$\hat{x}_k = \hat{x}^-_k + \mathbf{K}(z_k - \mathbf{H}_k \hat{x}^-_k) \quad (4)$$

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)\mathbf{P}^-_k \quad (5)$$

We track the position and velocity of each joint. To model the system kinematics, using the matrices $\mathbf{A}$ and $\mathbf{B}$, we assume that each joint is following a motion path governed by the usual equations of motion

$$x_t = x_{t-1} + dt\, u_{t-1} + \frac{1}{2} dt^2 a_{t-1} \quad (6)$$

If we define $\mathrm{x} = [\mathrm{x\_0, y\_0, ...}]^{\mathrm{T}}$, so that we have a vector of the x- and y-positions for M joints, and consider a model position as the state vector and velocity as input, we have a matrix equation that looks like

$$\hat{x}^-_{k+1} = I^{n \times n} \hat{x}_k + \Delta T^{m \times m} u_k + w_k \quad (7)$$

where dt is the timestep between frames and $\Delta \mathrm{T}$ is a matrix of timesteps. In the second case, with both position and velocity forming the state vector $[\mathrm{x, u}]$, our equation becomes

$$\hat{x}^-_{k+1} = \begin{bmatrix} I & \Delta T \\ \mathbf{0} & I \end{bmatrix} \hat{x}_k + w_k \quad (8)$$

In fact, for each of the three pose estimation methods (Optitrack [2], STAF [4], LightTrack [3]), we test four Kalman filters that describe different kinematic models:

1. Simple Linear Model: Position state vector $[\mathrm{x}]$, no input.
2. Velocity Input Model: Position state vector $[\mathrm{x}]$, velocity input vector $[\mathrm{u}]$
3. Constant Velocity Model: Position and Velocity state vector $[\mathrm{x, u}]^{\mathrm{T}}$, no input.
4. Acceleration Model: Position and velocity state vector $[\mathrm{x, u}]^{\mathrm{T}}$, no input, process covariance matrix $\mathbf{Q}$ contains acceleration information

4 Results: Speed and Performance

We investigate how the value of N, the number of frames tracked, affects the performance of the model. In the Figure, we see Multiple Object Tracking Accuracy (MOTA) total scores and computation speed when the three methods are run alone, and when they are run with tracking over N frames.

We can see that all three methods reach real-time tracking speed (30 FPS). For all three methods, there is a slight loss in MOTA performance with increasing N (8% and 14.5% for STAF [4] and LightTrack [3] respectively). This means that our module allows the use of some of these powerful pose tracking methods in real time, without sacrificing too much accuracy. We also transform a pose estimation method, OpenPose [2], into a pose tracking method.

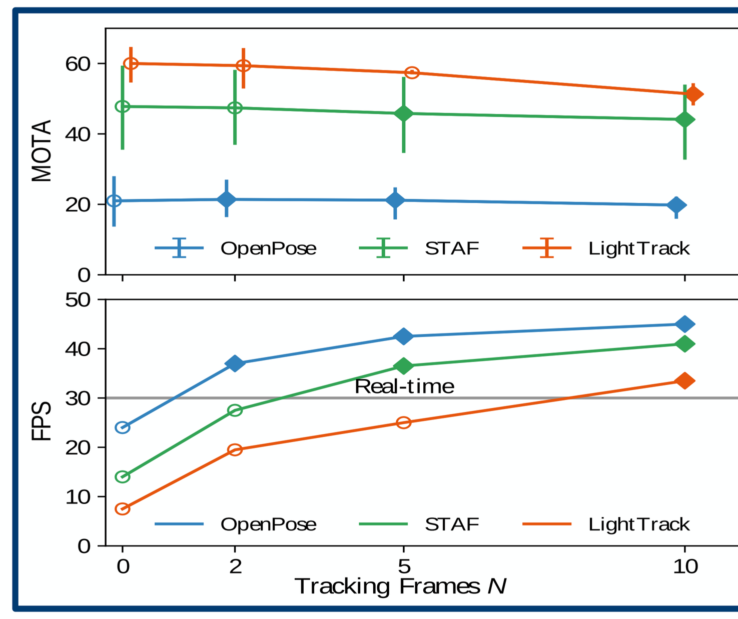
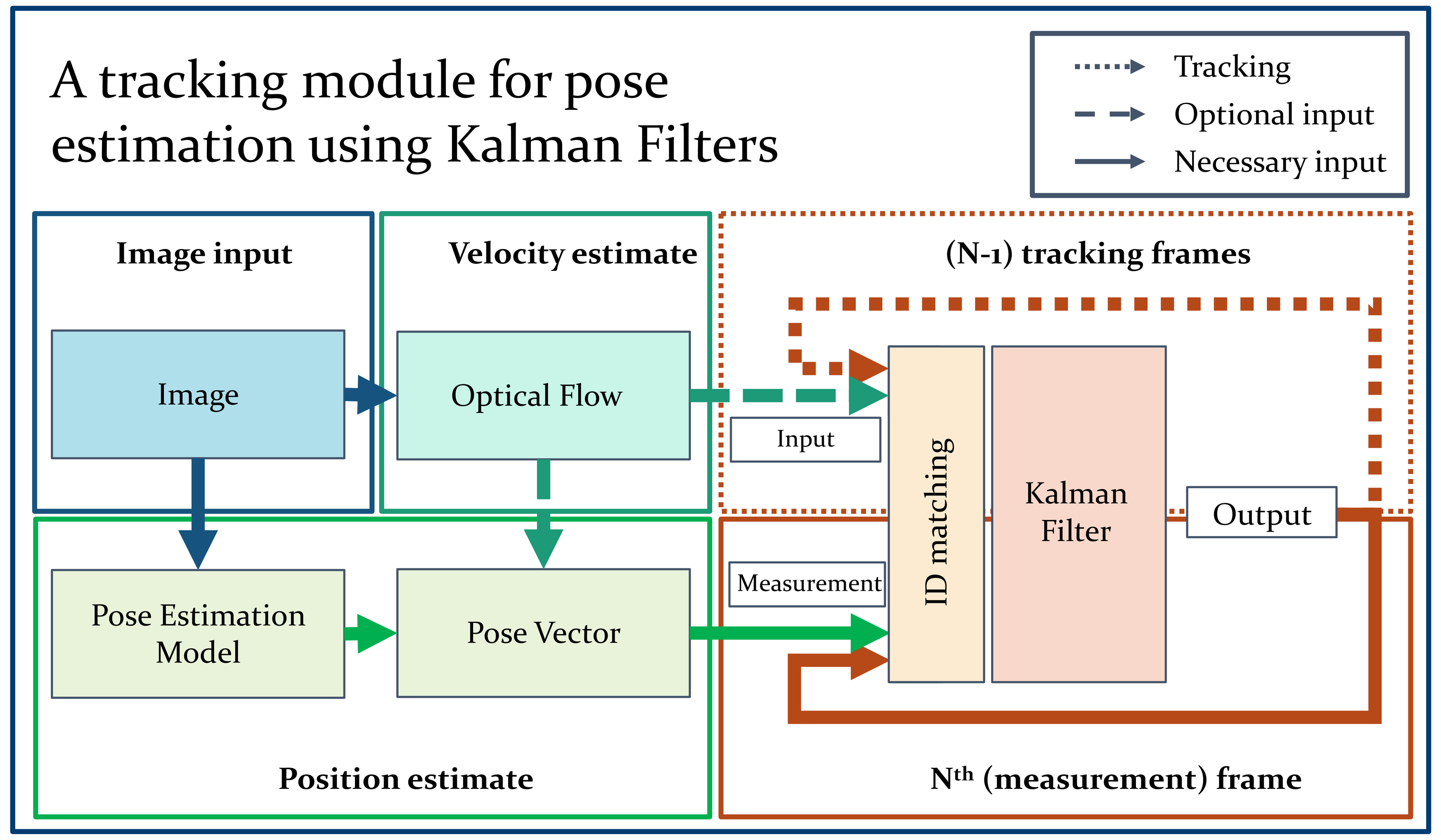


Figure 1: MOTA Totals and computational speed (FPS) with N = [0, 2, 5, 10] for each method. The diamond markers show where each method achieves a real-time speed of 30FPS or higher.

3 Method: Module Structure

1. Poses (a measurement) are taken from a pose estimation framework once every N frames
2. For (N-1) tracking frames, a measurement of position and velocity can be taken using an optical flow algorithm and is integrated with the estimate from the previous timestep



3. At every frame, ID matching is performed between current and previous pose estimates
4. The poses are then used to update the Kalman Filter, with any missing joint measurements replaced by a prediction from the Kalman Filter
5. The final estimate is a weighted average of the measurement and the state from the previous frame

6 Conclusion

We present a module that leverages both Data Assimilation and Machine Learning methods for human motion tracking. We can transform methods that perform *pose estimation* to *pose tracking* methods, or speed up existing *pose tracking* methods to real-time speeds. The different Kalman filters show how simple but considered changes to the internal structure of the system can tweak performance in different ways. In addition, these filters are computationally inexpensive, and can easily be run alongside a more expensive network to increase the speed at which a motion sequence can be analyzed.

Data Assimilation and Machine Learning methods are often similar enough in structure to complement each other. The method proposed here does not rely on any particularity of human motion estimation to be successful; thus there is no reason why it could not be applied to other areas. Several of the inaccuracies we encounter in human motion tracking are actually common in computer vision applications and are simply due to inconsistencies in the interpretation of images. Data Assimilation is a field founded on resolving and working with inconsistencies and uncertainties in datasets and therefore is an area that can make a valuable contribution to the computer vision field.

5 Results: Skeleton Tracking

Figure 2: Difference between unfiltered (top of each pair) and filtered (bottom of each pair) sequences from the PoseTrack dataset [1] for LightTrack [3] (top), STAF [4] (middle) and Openpose[2] (bottom). Note how for LightTrack and STAF, filtering allows more joints to be tracked throughout several frames. For Openpose, the poses are all identified, but the ID of each person is not kept constant, which is fixed by filtering.

References

[1] M. Andriluka, U. Iqbal, E. Ensafutdinov, L. Pishchulin, A. Milan, J. Gall, and S. B. PoseTrack: A benchmark for human pose estimation and tracking. In IEEE Conference on Computer Vision and Pattern Recognition, pages 5167–5176, 2018.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In IEEE Conference on Computer Vision and Pattern Recognition, pages 7291–7299, 2017.

[3] G. Ning and H. Huang. LightTrack: A Generic Framework for Online Top-Down Human Pose Tracking. arXiv preprintarXiv:1905.02822, 2019.

[4] Y. Raaj, H. Idrees, G. Hidalgo, and Y. Sheikh. Efficient Online Multi-Person 2D Pose Tracking with Recurrent Spatio-Temporal Affinity Fields. In IEEE Conference on Computer Vision & Pattern Recognition, pages 4620–4628, 2019.

Thank you for reading! For more information please read our paper or contact us at
{caterina.buizza11, t.fischer, y.demiris} @imperial.ac.uk

Please find more details of the work we do in the Personal Robotics Lab at:
www.imperial.ac.uk/personal-robotics/