Estimating infant upper extremities motion with an RGB-D camera and markerless deep neural network tracking: a validation study


Abstract—Quantitative biomarkers of infant motion may be predictive of the development of movement disorders. This study presents and validates a low cost, markerless motion tracking method for the estimation of upper body kinematics of infants from which proper biomarkers may be extracted. The method requires a single RGB-D camera, a 2D motion tracking software publicly available (DeepLabCut) and an algorithm generating 3D point coordinates from the 2D tracked points, dealing with missing data, originating from various sources, for estimating joint kinematics. The proposed method is validated using known point kinematics obtained from a doll, with size and shape of an infant, lying on a turntable rotating at 33½ rpm. Two camera image plane orientations are tested: parallel to the turntable motion plane and forming a 45° angle with respect to the motion plane. The latter enhances the occurrence of body parts occlusions during motion as expected in live infant motion recordings. The length of upper body segments, elbow and shoulder joint angles and the linear point velocity determined with the proposed method are evaluated against reference values obtained from the known motion of the turntable. The relevant Mean Absolute Errors (MAE) found indicate the margin of error to expect when processing live infant motion. The proposed method may be improved if enhanced hardware and tracking software are employed, therefore reducing the above-mentioned margin of error.

Clinical Relevance—The validation of the proposed method carried out in this study allows clinicians to select proper quantitative biomarkers obtained from infants upper body motion that may be useful for predicting movement disorders.

I. INTRODUCTION

It is important that deviations from the typical course of upper extremity motor skill development may be detected as early as possible so interventions can leverage the greater neuroplasticity that presents in the first year after birth. Clinician confidence in making an early diagnosis is impeded by the lack of specific biomarkers of the disorder and by the difficulty in recognizing patterns that clinically describe Cerebral Palsy (CP) [1]. Automated analysis of infant movements captured on 2D video has been under exploration for over a decade to aid in early identification of neuromotor impairments [2], [3]. Infants’ general movements (spontaneous, circular, “fidgety” movements) are characterized by small amplitude, moderate speed, and variable acceleration of all limbs in all directions [4].

3D video may provide a promising answer given its spatial resolution, depth information, accuracy and reliability, however technology cost and high computational overhead have been limiting factors [5].

Multicamera marker-based (MB) motion capture has been used successfully to distinguish differences in upper extremity movements in typically developing infants from those with neuromotor impairment [6], [7]. Recently, video-based markerless systems (ML) have been presented as a promising alternative to MB systems. Some ML systems are based on a deterministic approach including a kinematic model [8], others are based on a deep learning approach [9].

The current study aims at validating an innovative ML motion tracking approach based on a simplified instrumental setup and a well-established deep neural network tracking processing tool. Our system is suitable for the clinic or home, and uses a single, readily available device featuring RGB camera and depth sensor (RGB-D camera).

II. METHODS

The proposed ML method requires the use of an RGB-D camera and an RGB video tracking algorithm.

The camera used for this study is a RealSense D435 from Intel. The camera features a pre-calibrated RGB camera (native resolution: 1280x720, frame rate: ~30fps) and a depth sensor generating depth-coded images (native resolution: 640x480 frame rate: ~30fps). Each pixel of the depth image has an intensity proportional to the distance of the surfaces in the image from the camera. Depth images are generated by a stereo vision of two infrared sensors mounted on the device with the left sensor used as the main sensor.

If a region seen by the left sensor is not seen by the right sensor, then the depth of that region cannot be determined and the resulting depth image shows a “black area” in the corresponding region. RGB and depth images are pre-calibrated, however a residual misalignment between the two is present. The RGB video tracking algorithm used in this study is DeepLabCut, a ML motion tracking software for motion capture applications, originally developed for animal pose estimation*. DeepLabCut is based on transfer learning with deep neural

*This work was supported in part by the US Department of Health and Human Services (Administration for Community Living, NIDILRR) under Grant 90REGE0004 and by the “Fondo di ateneo per la ricerca - FAR 2019” (University fund for research).

D. Balta and A. Cereatti are with the Departments of Electronics and Telecommunications and Polito—Med Lab—Biomedical Engineering Lab, Politecnico di Torino, Torino, Italy (corresponding author phone: +39 3911264023; e-mail: diletta.balta@polito.it).

H.H. Kuo, J. Wang, M. Schladen and P.S. Lum are with the Department of Biomedical Engineering, The Catholic University of America, Washington, United States of America (e-mail: kuo@cua.edu, schladen@cua.edu, jingwangyu@gmail.com and lum@cua.edu).

I.G. Porco and U. Della Croce are with the Department of Biomedical Sciences, University of Sassari, Sassari, Italy (e-mail: Igorporco@uniss.it and dellacro@uniss.it).

* http://www.mackenziemathislab.org/deeplabcut
network using training data. The processing steps required by the proposed ML method are depicted in Fig. 1.

Figure 1. Block diagram of the proposed method

A. Acquisition of RGB and depth images and time alignment

Images acquired with the RGB sensor and those acquired with the depth sensor are aligned in time. Alignment in time was found to be necessary to compensate for irregular collection rates of the RGB images and the depth data. The timestamps provided by the acquisition software are used for this purpose. The three possible situations that could be encountered and the relevant countermeasures adopted are reported below:

A) The timestamp of an RGB image is closer to one or more RGB image timestamp than the closest depth image timestamp. Countermeasure: a gap of the proper number of frames is inserted in the sequence of depth frames;

B) The timestamp of a depth image is closer to one or more depth image timestamp than the closest RGB image timestamp. Countermeasure: a gap of the proper number of frames is inserted in the sequence of RGB frames;

C) The difference between the RGB and depth image timestamp is less than half the duration of the nominal sampling interval (<17ms). The two frames are considered time aligned.

All gaps generated are filled by applying a spline interpolation.

B. Tracking algorithm

RGB images are converted into video files (.avi) and fed to the DeepLabCut image processing tool which is based on transfer learning with deep neural network using training data. The software is trained to identify six points of interest (PoI) on infants’ upper body: left and right shoulders (LS and RS), elbows (LE and RE), and wrists (LW and RW). All PoIs are manually labeled on 10% of the video frames (selected by the software by using a k-means algorithm) to create a training set. After the training session, the tracking software provides PoI positions in all RGB frames, together with their relevant confidence level. When a PoI is occluded, the position is provided with a low confidence level. The training set was also set to 5% and 20% of the RGB frames for performance comparison purposes.

C. Association of the PoI depth coordinate to the PoI coordinates from the RGB images

The 3D position of PoIs tracked in the RGB images is obtained by exploiting the depth data. The location in the depth image corresponding to that of a tracked PoI in the time aligned RGB image is determined.

A number of issues could cause incorrect/or undefined PoI 3D positions (Fig.2):

- the estimated RGB location of a tracked PoI could fall over the “black area” in the corresponding depth image, therefore lacking depth information.
- the estimated PoI 3D positions could be corrupted by PoI occlusions. Since the tracking algorithm determines PoI locations exclusively from RGB information, the estimated location of a PoI could fall over a body segment covering the PoI in the RGB image (as when the head covers the shoulder). In such cases the estimate of the relevant depth coordinate is affected by an error equal to the distance, along the depth direction, between the surfaces of the two body parts. To compensate for this error, the prediction confidence level values provided by the tracking software are used. Such values report the estimated level of confidence (varying from 0 to 1) assigned by the tracking algorithm to every tracked point in every RGB image. A threshold on the confidence level was set to 0.6 and the depth values obtained for frames with a confidence level lower than 0.6 were not considered;
- an additional cause of errors in the estimation of the tracked PoI depth coordinate is due to the residual spatial misalignment between RGB and depth images. Such misalignment is responsible for errors in estimating depth coordinates when a tracked PoI is near a substantial depth discontinuity. To compensate for the consequences of this error, the following procedure was implemented: the first derivative of the PoI depth coordinate was calculated; when its value was higher than a threshold value set based on the physical limits of the subject motion speed then the relevant depth value was removed; all resulting depth coordinate gaps were then filled by applying a cubic spline interpolation, guaranteeing for first and second derivatives continuity.

Figure 2. Issues causing undefined 3D PoI positions: a) LS falling on the “black area” b) occlusion of the LS from the head c) residual spatial misalignment between RGB and depth images
**D Validation tests**

To evaluate the performance of the proposed method for correcting PoI depth coordinates, RGB-D camera recordings (about 5 seconds) of a doll positioned on a turntable rotating at 33⅓ rpm were carried out (Fig. 3). A recording captured the image plane parallel to the turntable rotation plane (0° acquisition); an additional recording captured the motion from a 45° angle from the horizontal plane to facilitate the observation of PoI occlusions (45° acquisition).

The upper arm (UA) segment length was defined as the 3D distance between the shoulder and the elbow PoIs, while the forearm (FA) segment length was defined as the 3D distance between the elbow and the wrist PoIs. UA and FA segment lengths were estimated in every instant of time for both body sides and compared to their reference values obtained with manual measures. Moreover, the angle of both elbows and shoulders were estimated. The elbow angle (EA) was defined as the angle between the FA segment and the UA segment. The shoulder angle (SA) was defined as the angle between the UA segment and the line connecting the two shoulders. The reference angles are calculated by manually identifying all PoIs in a static image.

Finally, the linear velocity of the RW is estimated from the RW 3D position in time. Its reference value is calculated by multiplying the nominal angular velocity of the turntable to the radius of the circle described by the RW trajectory.

The effect of the size of the training set on the estimates of the segment length during the 0° acquisition is shown in Fig. 4 where the relevant mean absolute error (MAE) are reported.

**III. Results**

The processing of the 0° and 45° 3D video acquisitions required the filling of a number of gaps due to the three possible sources previously listed and shown in Fig. 2. TABLE I reports the number of gap occurrences and their maximum duration for each tracked PoI.

**TABLE I.** NUMBER OF GAPS AND MAXIMUM DURATION FOR EACH CAUSE OF GAPS

<table>
<thead>
<tr>
<th>Issue causing gaps</th>
<th># of gaps [max duration (s)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time alignment</td>
<td>52 [0.33]  53 [0.33]</td>
</tr>
<tr>
<td>Oclusions + “black area”</td>
<td></td>
</tr>
<tr>
<td>LS</td>
<td>49 [0.43]  7 [0.07]</td>
</tr>
<tr>
<td>LE</td>
<td>2 [0.03]   2 [0.03]</td>
</tr>
<tr>
<td>LW</td>
<td>0           0</td>
</tr>
<tr>
<td>RS</td>
<td>58 [0.6]   1 [0.033]</td>
</tr>
<tr>
<td>RE</td>
<td>0           0</td>
</tr>
<tr>
<td>RW</td>
<td>0           0</td>
</tr>
</tbody>
</table>

**IV. Discussion**

The aim of the study is to preliminarily validate an innovative ML method for tracking the motion of infants. The method is based on the use of a single commercial RGB-D camera, to keep the experimental setup simple and affordable and a publicly available motion tracking software based on deep neural network. The validation of the method requires its application to known motions, such as that of a doll on a turntable, and the extraction of quantities related to parameters of potential clinical interest that can be obtained from the known motion and used as reference values, such as segment lengths, joint angles and point velocity. Designed for animal pose estimation, the selected tracking algorithm applied to the doll motion limited the errors in estimating the length of the upper body segments to less than 15 mm. A remarkable improvement is noticed when the size of the...
training set is increased from 5% of the total number of frames to 10%, while the improvement observed when the training set is increased from 10% to 20% appears to be limited, if present at all (Fig. 4), therefore indicating that utilizing a training set of about 10% of the total number of the video frames is an appropriate choice. This result supports the feasibility of our approach since the manual tracking of video frames necessary to form the training set is time consuming (about 30 minutes for a 6.6 seconds acquisition using the 20% of total frames option for the training set). As expected, the performance of the proposed method when the image plane is parallel to the plane of motion (0°) is remarkably better than when the image plane forms a 45° angle with respect to the plane of motion. This result provides a crucial experimental setup design criteria for infant motion recordings. The errors in determining the segment length are influenced by the recorded gaps and the limited capability of the tracking algorithm to track the PoIs when occlusions occur, or more generally, when the PoI is not clearly identified. In our validation, the errors in estimating the velocity of the wrist are generally less affected by the above-mentioned issues, making the reliability of clinical parameters extracted from the wrist trajectory quite reliable. In general, the errors shown by the method are due to a number of reasons, some directly related to the limitations of the hardware, some to the limitations of the tracking algorithm, some to the camera orientation and some to the intrinsic nature of tracking body parts using points. More specifically:

- Hardware limitations: irregular frame rates of both RGB and depth sensors cause gaps in RGB and depth images. In addition, the “black areas” detected in the depth images are responsible for additional gaps in the sequence of depth images. Even if all gaps are filled with cubic interpolation splines, when their duration extends over a few frames, the resulting 3D position estimates suffer from important approximations;

- Tracking algorithm limitations: the selected algorithm for tracking the motion recorded in this study was not designed to track infants motion but rather live animals motion and there is a limited number of settings that the user can adjust to improve its performance. Moreover, being based solely on RGB images, the algorithm does not take advantage of the presence of the depth information in the tracking process;

- Camera orientation limitations: the tests performed highlight the importance of the camera orientation with respect to the plane on which most of the motion occurs, when present, since this method is developed to track the motion of the upper body of infants seating in a baby chair, it is advisable to position the camera frontally to limit body segment occlusions;

- Tracking body parts limitations: ML motion tracking methods applied to humans (or animals) often track the motion of body surface areas and record the motion of a single point, frequently associated with internal joint centers, resulting in errors caused by the three dimensional nature of the human (and animal) joints.

Some of the above-mentioned limitations may be dealt with by using an RGB-D camera with an improved depth sensor, properly positioned in consideration of the expected motion to be recorded and a tracking algorithm developed for 3D images, incorporating the model of joint centers as internal points.

Another limiting factor of the use of the selected tracking algorithm is the time required to manually track even 10% of the RGB images. A customized tracking algorithm could reduce the burden of the generation of the training set. However, even with the current limitations, the results obtained in this study may provide useful elements for a foreseeable application to recordings of infant motion and the extraction of biomarkers of clinical interest for the early detection of the development of movement disorders.

V. Conclusion

The aim of this work is to test the effectiveness of the proposed method based on ML recordings from a single RGB-D camera to estimate infant’s motion. Marker-based protocols are still the gold standard in clinical practice, but there is certainly a need for simpler and faster assessment protocols, especially on infants. Future studies will be devoted to validating the proposed protocol on infants for testing its use in clinical practice.

REFERENCES


