STORM-Net: Simple and Timely Optode Registration Method for Functional Near-Infrared Spectroscopy (fNIRS)

Yotam Erel
School of Computer Science
Tel Aviv University
Tel-Aviv, Israel
erelyotam@gmail.com

Sagi Jaffe-Dax
Department of Psychology
Princeton University
Princeton, NJ, USA
jaffedax@princeton.edu

Yaara Yeshurun
School of Psychological Sciences
and Sagol School of Neuroscience
Tel Aviv University
Tel Aviv, Israel
yaaray@tauex.tau.ac.il

Amit H. Bermano
School of Computer Science
Tel Aviv University
Tel-Aviv, Israel
amberman@tauex.tau.ac.il

Abstract

Significance: We propose a robust video-based method for estimating the positions of fNIRS optodes on the scalp.

Aim: Calibrating the location of optodes relative to a subject's scalp is an important step in acquisition of reliable neuroimaging data, and is a relatively open problem when dealing with developmental populations. Existing methods pose various motion constraints, require expert annotation and are only applicable in laboratory conditions. A quick and robust framework to deal with these issues is required.

Approach: Using a variety of novel computer-vision technologies, we implement a fully-automatic appearance-based method that estimates the registration parameters from a raw video of the subject. We validate our method on 10 adult subjects and prove its usability with infants as well.

Results: We compare our method with the golden standard 3D digitizer, and to other photogrammetry based approaches. We show it achieves state-of-the-art results. Our method is implemented as a freely available open-source toolbox at `https://github.com/yoterel/STORM`.

Conclusions: Our method allows to calibrate the fNIRS system in a simple way, with unprecedented speed and accuracy. Fast calibration facilitates more spatially precise neuroimaging with developmental and clinical populations even in unconventional environments.

Keywords: functional near-infrared spectroscopy; developmental neuroimaging; spatial co-registration; photogrammetry; convolutional neural network

1 Introduction

Functional Near-Infrared Spectroscopy (fNIRS) is an emerging neuroimaging technique that is becoming the preferred method for acquisition of cortical activity in developmental and clinical populations [9, 16]. Its wide usage in developmental research is due to three main advantages. First, fNIRS

Preprint. Under review.

provides better comfort for both the participants and the experimenter compared to fMRI. Second, it allows higher spatial separation compared to EEG. Third, it is less susceptible to participant's motion during the study compared to both fMRI and EEG. These advantages makes fNIRS the ideal method to measure the brain response under relatively ecological settings. Nevertheless, the level of confidence and acceptance of this neuroimaging technique is limited due to the lack of a standard method for inferring the underlying cortical region which provides the signal for a specific channel. There are two challenges in pursuing this standard method. First, the need for an exact estimation of the scalp placement of the probes. Second, the requirement for a standard age-dependent cortical template that matches each scalp placement with a specified underlying cortical structure (e.g., gyrus or Broadmann area). Previous studies have already estimated the probable underlying cortical structure for scalp locations [18, 11]. Recent studies have also established the age-dependent template that should be used [6] and even corrected for head size variations [21]. However, the first challenge has been relatively overlooked, or solved only by constraining the portability of the system and reducing its usability with developmental or clinical populations. These constraints include: the requirement for attentive motionless subjects [7], a metal-free environment [1, 15], a setup that constitute a contraption unfit for an infant [17], a need for specific illumination (mono-chromatic background, well lit area, strongly colored cap, etc.) [8], post-experiment calibration [3] and intensive expert manual labor [13, 4].

In general, contemporary methods for estimating optode locations on the scalp are either hardware based or photogrammetry based. In the hardware based approaches, the most popular device types are electromagnetic localizers (e.g. [5]). These systems offer great accuracy but are limited by interference from metal objects around the participants, and require them to remain motionless for up to several minutes during calibration. Early developmental and numerous clinical populations fail to meet this requirement. Moreover, the system itself is costly, and offers limited portability. Similarly, other hardware based solutions usually suffer from the same disadvantages, as above [17].

Photogrammetry-based approaches use optical devices such as cameras to acquire (possibly multiple) 2D representations of the system, and calibrate it relative to some known ground truth. Since they use the visual appearance of the cap as their source of information, their advantages include portability and simplicity of hardware requirements, and relieving the main constraints posed by the 3D digitizer as described above. The method used by Lloyd-Fox et al. [13] is an example of using images to calibrate the fNIRS, but they require expert manual annotators to mark positions of fiducials. Hu et al. [8] performed registration with the SfM algorithm for 3D reconstruction of a scene based on multiple images from different angels. However, their method still requires manually selecting the sensors' location on the 3D model after reconstruction. Still, this offers much greater convenience and does not require any manual measurements to be performed. Furthermore, the algorithm heavily relies on good selection of matching feature vectors extracted from the set of images taken and on the SIFT algorithm to find them, which can yield inferior results under various lighting and filming conditions. This is especially noticeable for footage lacking a feature rich environment [22], such as the black cap used by Hu et al. [8]. Jaffe-Dax et al. [10] used a similar approach, but overcome the manual annotation with a template matching algorithm thus achieving a fully automatic solution. To counter the feature matching problem mentioned above, the authors printed a feature rich pattern on the cap. However, since the SfM algorithm takes a significant time to complete, the fully automated system fail to reach its full potential: Infants may lose attention during this time, move abruptly, thus preventing the required images for the reconstruction to be successful. In addition, this method can only be used as a post-processing step as opposed to a possibly more desired on-the-fly configuration. In both methods described, the profile of image capturing required is extremely difficult to achieve with developmental population, since they require the participants' focus to be away from the moving camera through most of its path. Drawing their attention to some other place while filming is challenging even for experienced and proficient neuro-scientists.

Despite their convenience, state-of-the-art works still suffer from sensitivity to changing illumination conditions, relatively complex acquisition instructions and most importantly, a lengthy and cumbersome calibration process.

In this paper, we suggest a new calibration method. In the heart of our solution lies STORM-Net, a Convolutional Neural Network we developed which predicts cap orientation from raw video. Our calibration process is performed in under 10 seconds, allows immediate calibration results, using nothing but a commercial low budget smart phone or camera and is suitable for non-laboratory conditions due to its robustness. Moreover, we conducted an experiment in order to validate our method's accuracy, by comparing our calibration results to the popular 3D digitizer approach[5].

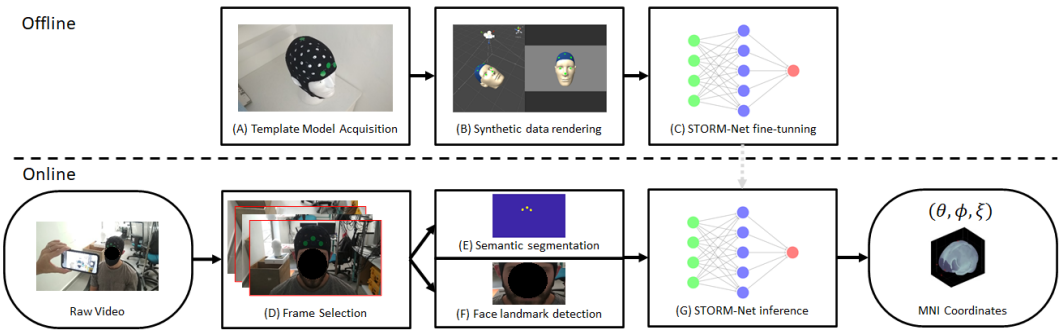

Figure 1: STORM-Net system overview. Given a new cap, the offline step is performed once (top), while the online step is performed per subject in real-time (bottom). (A) Cap and optode locations are acquired (e.g. by using a digitizer [5], or by the method introduced by Jaffe-Dax et al. [10]), (B) Synthetic data is automatically generated using a custom renderer (left: the virtual scene, right: projection of virtual scene to camera plane), (C) A STORM-Net model (tasked with estimating the cap-head registration using the three Euler angles) is trained on the synthetic data, either from scratch, or as a refinement, starting from a pre-trained model, (D) 10 frames are selected from a raw video, (E,F) Facial landmarks and cap fiducials are extracted from the video using a semantic-segmentation network and a face extraction model, (G) STORM-Net predicts the registration parameters and MNI coordinates of all optodes are retrieved.

2 Method

Our video-based approach is shown in Figure 1. Given a new cap, it is first registered in an offline step (i.e. without any human subjects present) for improved accuracy. This produces a network that is fine-tuned to the specific cap, and is utilized in real-time to calibrate the sensors' positions on a human subject. Taking a deep learning approach, we aim at a concise solution to the problem in an end-to-end manner. In other words, a single network that starts with a video and performs calibration, all in one pass. This is desired when possible since it yields real-time performance and gives the network a holistic understanding of the problem (as opposed to exposing different networks to different steps in the process), thus is much less prone to errors. However, this approach poses a significant challenge: A suitably large data-set must be available since we need to estimate a much more sophisticated function over the input (which also implies larger training and prediction times). In our case, producing such data-set is not feasible at this scale. Realistically, one is able to create only a handful of videos, since their capture and annotation (i.e. through digitizer measurements) are slow and tedious. For this reason, we chose a hybrid approach: our video processing pipeline is split into several blocks performing separate logical tasks. To decrease the amount of required data, some of the blocks use well-explored technologies that reduce the raw data (i.e. the video clip) into much smaller dimensions (2D "landmarks") so annotation becomes significantly easier, and can be performed automatically. We still maintain the holistic view approach for the core network (STORM-Net) by exposing it solely to these landmarks which are theoretically sufficient to conclude the calibration parameters using prior geometric knowledge of the system (Section 2.1.1). Furthermore, since using landmarks decouples the appearance and geometry of the cap (i.e., the whole image vs. 2D landmarks of particular 3D points) we can use synthetic data instead of actual videos, which dramatically improves the usability and adaptability of the method. Lastly, the modularity of the pipeline makes its predictions more explainable; if some prediction carries a large error, it is easier to detect which block failed and how to fix it compared to a one big black-box neural network solution.

In the offline step (Figure 1, top), the sensor locations are measured in 3D in a controlled environment using an external digitizer device [5], or any other 3D coordinate acquisition method (e.g. Jaffe-Dax et al. [10]) on a mannequin head (Figure 1, (A)). Then, a user-defined subset of the sensor locations is used to produce synthetic 2D images, simulating a short video session (B). We use this data to train a rotation-estimation neural network (STORM-Net) that predicts the orientation of the entire cap from these 2D locations (C).

In the online step (Figure 1, bottom), the user takes a short video of a participant wearing the pre-registered cap to calibrate sensor positions in real-time. The 2D locations of the aforementioned

user-defined subset are then automatically extracted from the video (Figure 1, D,E,F), and are fed into STORM-Net (G). By applying the transformation predicted by STORM-Net to the template model, the final output is the 3D location of all sensors, represented in a desired coordinate system (such as MNI coordinates).

2.1 Offline Step

When using a new cap for the first time, the offline step is required. The user prepares an "ideal" environment where the cap is mounted on an anatomically correct mannequin head while being well aligned and positioned relative to it. The user then manually marks the positions of 4 or more specific locations (in all our experiments, we have used $\{"Fp1", "Fpz", "Fp2", "Cz"\}$) on the cap by placing colored stickers on them. The 3D locations of these cap landmarks, 3 facial landmarks (the two eyes and the nose tip), along with all other cap sensor positions, are then measured (e.g. by using an external digitizer device [5]). These 3D locations are referred to as the *template model* throughout the paper. The template model is fed into a renderer, which produces synthetic sequences of frames imitating the geometry and kinematics of a head wearing the cap in real videos. Each such sequence is a matrix of size $10 \times 2N$, representing a video with 10 frames where each frame has N (2D) positions of the cap and facial landmarks mentioned earlier ($N = 7$ in our case). The rendered sequence approximates the real video properties of the expected input in the online step. The produced samples are then used to train STORM-Net. Once trained, STORM-Net is able to predict the orientation of the cap relative to the scalp in all 3 axes in real-time. The entire training process takes under 6 GPU hours (but still possible without one). The separation of this offline step from the actual measurements allows a very short calibration time (10 seconds) during a live session, yet has the extra benefit of increased accuracy. We found that instead of supplying a pre-trained network to perform this task, fine-tuning it on user supplied synthetic data (created from the template model) improves results dramatically. See the supplementary material for comparison.

2.1.1 Cap Preparation

As previously mentioned, we manually marked the positions of 4 cap landmarks, $\{"Fp1", "Fpz", "Fp2", "Cz"\}$, using stickers. It is noteworthy that using these locations, or any standard anatomical landmark for that matter, is not necessary. Any sufficiently informative set of 3D positions is possible. Sufficiently informative in this case means that the matrix representing the set is at least of rank 6 (i.e. all 6 parameters of a rigid body in 3D can be deduced). The more landmarks used, the better the accuracy achieved (with diminishing effect), but measuring them makes the setup more cumbersome. Figure 3 visualizes the 3D to 2D landmark projection process. In our experiments, we used green colored stickers to mark the landmarks but this is not essential - their 2D positions are faithfully extracted by a semantic-segmentation network in the online step which was trained on merely 300 annotated photos, with very high reliability. A detailed description of the training process can be found in the supplementary material.

2.1.2 Synthetic Data Generation

One of the biggest drawbacks of data-driven approaches in general, and neural network based learning in particular, is the need for a large body of example data to train on. We overcome this obstacle by introducing a framework that is based *solely* on synthetic data. To enable this, we train the network on landmark positions, instead of raw images, since those are easily imitated by synthetic means. Obviously, the prediction accuracy of STORM-Net on real videos is highly correlated with the ability to model and simulate them as best as possible. We use our renderer to perform this task, examples of which can be seen in Figure 2.

Using our renderer and the coordinates measured according the process described in Section 2.1.1, we generate sequences of 2D locations that emulate the capturing process described in Section 2.2. Since the human measuring process can never be ideally accurate, the most crucial part of the data synthesis process is correctly expressing the distribution of possible human input. The first deviation from an ideal scenario is that the camera never follows the exact required path, depicted in Figure 3. We handle this using 3 techniques. Firstly, we add random noise to the camera's position and orientation during simulation (which we call the "Shaky Cam Effect"). Additionally, we add range deviations by increasing or decreasing the camera's distance from the scene linearly with time. This is because we found that while taking the video it is often hard to keep the distance from the participant constant.

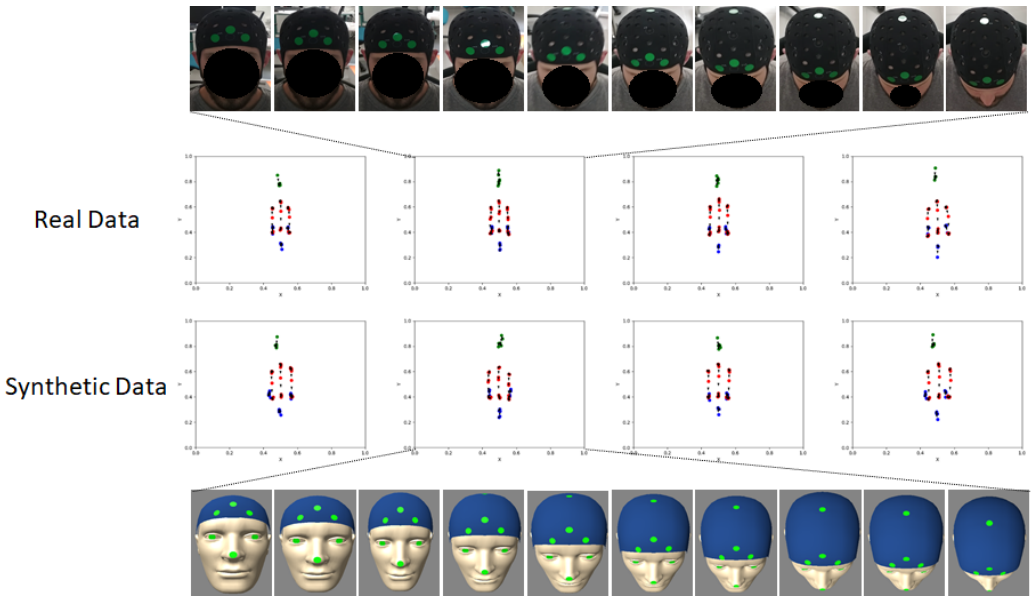

Figure 2: A visualization of real data (top) and similar synthetic data (bottom) extracted from different videos and rendering sessions respectively (each column). The points in the plots represent the 2D positions of the landmarks across a whole sequence. They are colored blue for facial landmarks, red for cap landmarks, and green shows explicitly the location of "Cz" for reference point. The data is centered and normalized. Notice how the synthetic data closely mimics the distribution of the real data and its noise sources. We used these graphs to aid with modeling the noise sources during scene creation in the renderer.

Lastly, to mitigate the effects of the participants' location being outside the center of the frame, we always center all landmarks using their center of mass as a post-rendering routine (i.e. after the data is rendered but before it is fed to the network).

Another type of noise we take into account is partial occlusion. This happens because of the nature of the defined path, where the cap and face position relative to the camera occlude certain locations (e.g. the eyes are not visible when filming from above, which is the final position of the camera in the defined path). To handle this, we modeled a 3D representation of the face surface such that occlusions occur automatically by rendering the scene. We also artificially remove a random number of stickers from each frame of each data-point during training, inspired by popular drop-out techniques [20]. Hidden landmarks are placed at $(x, y) = (0, 0)$, assuming the neural-network learns to ignore all such points given sufficient examples.

Lastly, selecting frames as discussed in 2.2.2 introduces temporal inconsistencies, such as non-smoothness, or even non-monotonic behavior, e.g., the cameraman momentarily shifts the camera in the opposite direction of the regular path. We handle such errors by randomly shuffling pairs of consecutive frames in the data-points. To observe the benefits of this augmentation has on results, see supplementary material.

2.2 Online Step

In this step, subjects are fitted with the same cap used in the offline step (with the stickers). The user then captures them on video using a camera (we used a low-budget smartphone) that moves along a predetermined path for about 10 seconds. See Figure 3 for an exact description of the camera path. This video is then processed in the following way: we automatically select 10 frames from the video and feed them into an off-the-shelf facial landmark detector (based on a neural network) and our pre-trained semantic-segmentation neural network (which was trained to segment fiducials out of frames - green stickers in our case). These neural networks were designed to output 2D locations of the facial and cap landmarks mentioned above for every frame. Together their output consists of what

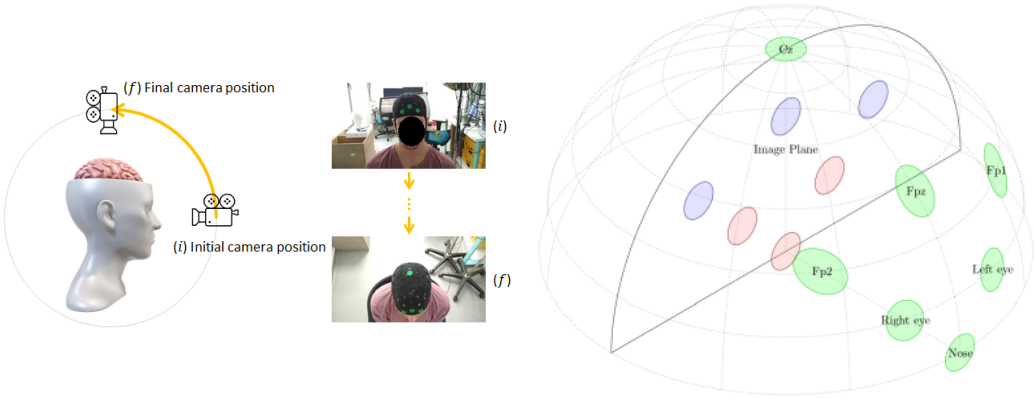

Figure 3: Left: The camera path designed for our system. Ideally, the initial camera position is exactly in front of the subject (i), and in a 10 second time frame it follows a circular path, reaching the top of the head (f). Right: An illustration of projecting facial and cap landmarks to the camera's image plane. Red projections indicate facial landmarks whereas blue projections indicate cap stickers (fiducials). Using a sequence of images taken from different directions, the data encodes sufficient information enabling the neural network to learn the orientation relative to scalp. Note that landmarks that are perpendicular to the image plane (such as "Cz" in the image) or landmarks occluded by the head are not visible in the projected image, causing loss of information that occurs during the capturing process, emphasizing the need for multiple images to be taken.

is needed by STORM-Net to predict rotation parameters. We obtain these, and applied them to the template model to get calibrated positions of all sensors relative to brain coordinates.

2.2.1 Video Capture

Our method uses raw video data to extract 3D information. Thus, it is essential this video includes useful data on the 3D relationships between the marked locations. Useful, in this case, means we are able to deduce the cap orientation and position relative to the scalp. As mentioned in (2.1.1), one way to make this possible is adding enough landmarks. Another useful way is using multiple images (video) of the same scene. To this end, we designed a camera path which follows a circular curve, ensuring these relationships are indeed captured when taking the video, while reducing interference with participant attention (which is imperative with the infant study group). Additionally, this path is easy to follow with a free hand held camera, without dedicated equipment, as demonstrated in Figure 3. In our case, videos were all taken using a low-budget smartphone with 1920×1080 resolution and 30fps.

2.2.2 Selecting the Frames

We found that using a somewhat naive selection generally suffices. We start by picking 10 frames evenly spaced from the video and naming them "base sites". We observed that some specific frames are extremely blurry due to sudden movements of the hand held camera or the subject's head. Additionally, the camera we used was of a rolling-shutter type and relatively of low performance. To mitigate this, we measured the amount of blurriness in a temporally local environment near every base site using the variance of the Laplacian operator [14], and selected the best candidate from every such environment (10 such selections are made). Nevertheless, this approach may fail if the video was taken based on a camera path significantly different than required.

2.2.3 Extracting Facial Landmarks

The participant face is identified using an off-the-shelf facial landmark detector [12] as can be seen in Figure 4. We define a polygon around each eye's landmarks, and take their average position as the eye's "location". For the tip of the nose, we use one specific landmark. If a face is not detected in some frame, then all coordinates of the facial landmarks are set to $(x, y) = (0, 0)$ (for that frame). Note that facial landmarks are identifiable and ordered, i.e. given an output from the facial detector

we can always uniquely tag each of the 3 (x, y) pairs with either "Left Eye", "Right Eye", or "Nose Tip". This is important when estimating rotation parameters, as explained in Section 2.2.5.

2.2.4 Extracting Cap Fiducials

Each frame selected is passed through a pre-trained semantic-segmentation neural network inspired by "capnet" [10] (see supplementary for architecture details and training parameters). The raw frame is then transformed to a binary mask where true valued pixels belong to the green stickers, marking specific cap locations. This binary mask is further processed into discrete 2D locations by locating their center of mass. To do this, we used the *findContours* method [2] to detect the contours of the blobs present in the binary mask and filter out blobs that were small or located very far from the center region of the image. These thresholds were determined experimentally, but were observed to be robust in various scenarios. Finally, the center of the remaining 4 largest contours are reported as the cap landmark positions. If there are less than 4 remaining, the rest are set to $x, y = [0, 0]$. Note that the contours found from the binary mask have no defined order. The 4 (x, y) pairs defining the cap landmark locations are thus an order-less set, and must be treated as such. Note that although the segmentation network could have been adjusted to output an ordered set (i.e. instance segmentation), this would require a much more extensive data-set to produce accurate results and is relatively straight-froward to do based on the image appearance rather than implicitly understanding it.

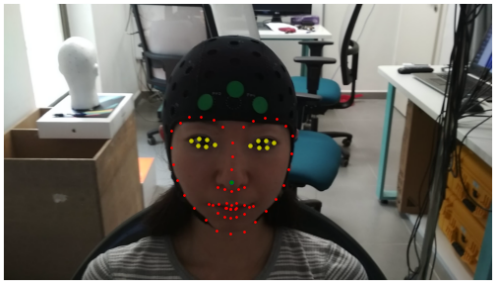

Figure 4: A frame with 68 facial landmarks extracted using an off-the-shelf extractor [12]. The eye landmarks (marked in yellow) are averaged to obtain the final eyes position and the tip of the nose (marked in green) is used as-is. Note that this landmark extractor is selected for its simplicity and effectiveness in our conditions, but can be seamlessly replaced by another.

2.2.5 Estimating the Rotation Parameters

For this block, we arranged the landmarks location data into an $F \times N$ matrix, where F is the number of frames (10 in our case), N is the number of landmark coordinates ($7 \cdot 2 = 14$ in our case). Note that the first 3 pairs in every row represents an ordered set corresponding to the left eye, right eye and nose tip respectively. The other 4 pairs of every row are order-less, each pair representing one cap landmark point (in our case from the set {"Fp1", "Fpz", "Fp2", "Cz"}). We feed this data to STORM-Net, a 1D Convolutional Neural Network (CNN) that outputs 3 Euler angels $\{(\theta, \phi, \xi) \in \mathbb{R}^3\}$. These are the rotation parameters that indicate how should the template model be rotated to fit the observed data. As mentioned in Section 2.1.2, STORM-Net was trained using synthetic data produced by our renderer (see supplementary for architecture details and training parameters). To stay true to the real videos, we randomly shuffle the order-less pairs before they are given to the network. During our experiments, we have noticed that shuffling the facial landmarks pairs yields significantly worse result, even when the cap landmark points are left intact. This is a good indicator that the network learns a good internal representation of the 3D rotation parameters, since the facial landmarks and their identity are imperative to anchor and deduce the cap orientation relative to the scalp.

2.3 Experiments

We conducted human experiments to validate our method. All experimental procedures were approved by the Tel-Aviv University Institutional Review Board for Human Subjects. All subjects provided written informed consent.

2.3.1 Participants

10 subjects participated in this study (8 males, 2 females, aged $29.8Y \pm 2.73Y$ $(mean \pm STD)$. Their heads sizes were $X_1 = 35.7cm \pm 1.54cm, X_2 = 39.9cm \pm 1.8cm$ $(mean \pm STD)$ where X_1 is the distance from the Inion to the Nasion along the scalp, and X_2 is the distance between

Pre-auricular points along the scalp passing through the midway point of X_1. Notice our selection of subjects with "medium" sized heads was made on purpose, so the cap we used fit well. This does not imply our method can't work with other head sizes - this would simply require repeating the offline step as shown in Figure 5 for caps that better fit different head sizes.

2.3.2 Experimental Design

Participants were fit with the cap which was well-positioned on their scalp (using standard positioning methods, without any external means). Two videos were taken followed by two digitizer measurements of certain landmarks while participants were wearing the cap - these were used to ascertain the intra-method reliability of the two methods and their inter-method validity. For the videos, this was done by feeding them to STORM-Net and comparing results. For the digitizer measurements, this was done using rigid body registration (Section 2.3.3). The cap was then removed from the scalp and re-positioned. Another video and another digitizer measurement for the landmarks were taken and were used to estimate the shift in positions of the channels by the two methods.

2.3.3 Rigid Body Registration

Given a rigid body model for the cap and two set of optode measurements $X_1, X_2 \in R^{nx3}$; where n is the number of optodes and X_2 was rotated and translated in some unknown manner, we use SVD decomposition [19] to obtain optimally aligned sets in the least squares sense. This was used to compute optode positions and shifts as measured by the digitizer for comparisons.

2.4 Implementation

Our software stack is written in python 3, and is freely available on Github:

`https://github.com/yoterel/STORM`

The toolbox includes a GUI (graphical user interface) and a CLI (command line interface). The CLI allows a fully automated real-time end-to-end calibration, while the GUI offers more subtle control over the process. The GUI includes a 3D model viewer (for visualizing the template model), a video annotation tool (users can manually or automatically perform annotation), the synthetic data renderer, and automatic training scripts for all described neural networks in this paper. Most importantly, it also allows the calibration of a given video file.

3 Results

3.1 Neural Network, Validated Through the Standard 3D Digitizer

Overall our method's accuracy was comparable to the standard 3D digitizer method. Across all subjects ($n = 10$), the distances between the positions of the same channel as estimated by the two methods were similar to the distances between the positions of the same channel as measured by the digitizer twice (inter-method validity of $3.10mm \pm 1.87mm$ compared to intra-method reliability of $2.56mm \pm 1.14mm$; $mean \pm STD, t_9 = 0.88, P = 0.40$). The intra-method reliability of our method ($2.73mm \pm 2.25mm$; $mean \pm STD$) is somewhat lower than the digitizer ($t_9 = 0.17, P = 0.865$). However, the error of our method is still an order of magnitude smaller than the typical distance between optodes ($3cm$). All the subjects were fit with the same cap, and we have not observed any bias in prediction accuracy with regards to certain head sizes as shown in Figure 5.

3.2 Neural Network Cap Shift Detection

After placing the cap and performing validation measurements twice, as described in the previous section, the cap was removed and re-positioned on the subjects head, and new measurements were performed (Using a digitizer and our method). As expected, this re-positioning process introduced some shift in optode positions. We found a high correlation between the measured shift in optode positions using the digitizer and the measured shift using our method. Since we use a rigid model registration it is sufficient to describe such shifts using the rotational parameters θ, ϕ, ξ which represent the rotation around the X axis, Y axis and Z axis respectively. We obtained Pearson's

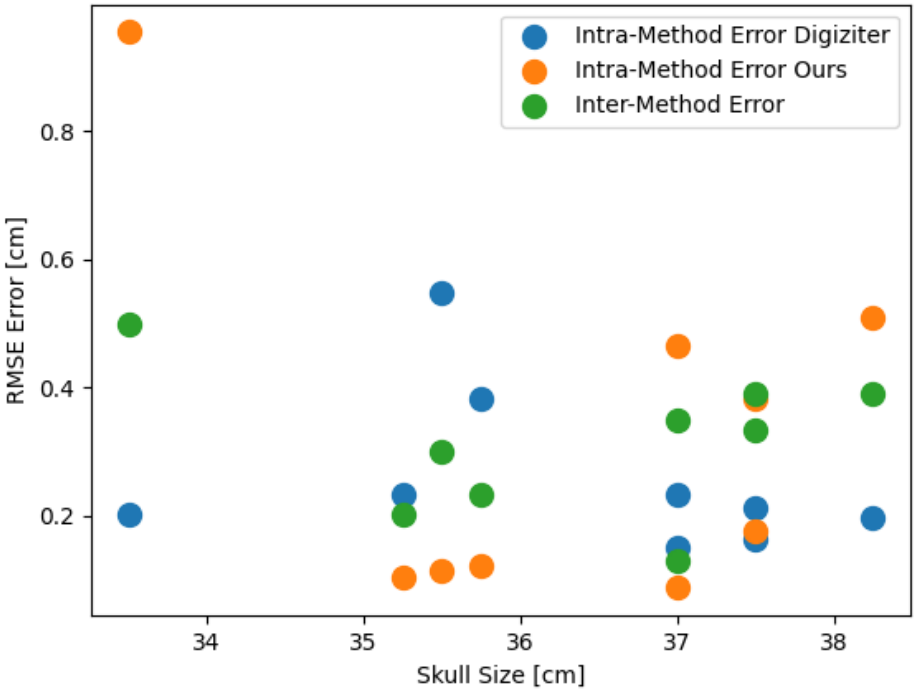

Figure 5: RMSE error as a function of skull size. The RMSE error for various skull sizes of different subjects when using the digitizer twice (Intra-Method Digitzier), our method twice (Intra-Method Ours), and between the digitizer and our method (Inter-Method). Skull sizes were measured according to the description in Section 2.3.1. Notice that our method does not produce significantly different errors across the subjects, except for one, who had an extremely smaller skull. For this subject, results might not be accurate, due to inadequate fit of our cap which was designed specifically for medium sized skulls. Nevertheless, this verifies that our method is robust to the size of the patients head as long as the cap fits. Otherwise, a new template model must be created offline, for a more suitable cap.

coefficients of $0.585, 0.670, 0.794$ for θ, ϕ, ξ with $P = 0.098, 0.048, 0.01$, respectively, across all subjects, which indicate a positive close to linear relationship between shifts detected by the digitizer and shifts detected by our method. Importantly, our method is accurate when predicting the rotation around the Z axis (perpendicular to the ground). Positioning of the cap by skilled practitioners starts by placing "Cz" in the halfway between the Nasion and Inion and halfway between the ears and only then the rest of the cap is properly placed. This manual technique yields much better human precision with regards to rotations around the X and Y axis since fixing "Cz" to the scalp at an anatomically nearly precise position "constrains" the possible cap movements to Z rotations.

3.3 Automatic Channel Estimation in Infants

We performed a qualitative evaluation of our method on infants using a different template model, and observed results which were consistent with our speculations, even for very large rotations of the cap relative to the scalp as shown in Figure 6. When the cap is positioned with an angle relative to the scalp, STORM-Net still easily predicts the rotational parameters for each video. These parameters can be used to register the cap and obtain the 3D positions of all optodes using a template model of the cap.

Table 1: Comparison of photogrammetry based approaches

	Hu et al. [8]	Jaffe-Dax et al. [10]	**Ours**
Inter-method error (adults, $AVG \pm STD$	$6.66mm \pm 3.30mm$	$3.40mm \pm 0.90mm$	$3.10mm \pm 1.87mm$
Acquisition Time **Processing Time**	~ 1 minute > 10 minutes	~ 1 minute ~ 10 minutes	~ 5 seconds < 0.1 seconds
Automation of calibration step	semi-automatic	fully-automatic	fully-automatic
Camera specifications	8MP resolution per image	2.1MP resolution per image, 240 fps	2.1MP resolution per image 30 fps.
Environmental constraints	mono-colored walls or cloth/curtains covering the background.	None	None
Sensitivity to cap appearance	High	Medium	Low

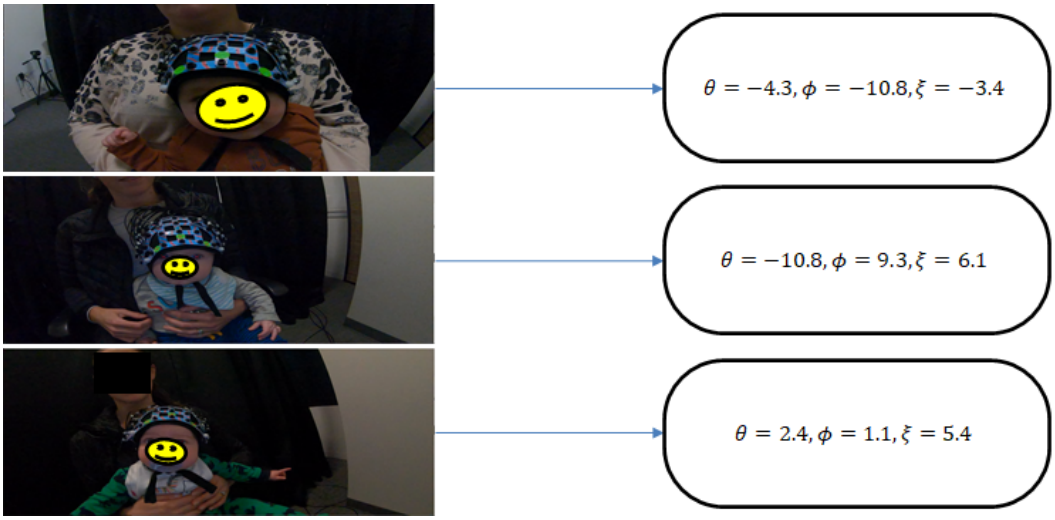

Figure 6: Qualitative evaluations of STORM on infants: The figure shows the 3 rotational parameters obtained using STORM from various infants videos. We observed that the parameters conform to our speculations in every axis across all videos we tested. For example, the first 2 videos (top 2 images) got negative and positive values for ϕ, which is expected for rotations around the Y axis (back to front). Notice the parameters are given in a right-handed coordinate system (i.e. angels can be interpreted as rotations around their respective axes, using the right hand rule).

3.4 Comparison to Previous Works

We compared our method to other photogrammetry based approaches used for registration of the optodes presented by Jaffe-Dax et al. [10] and Hu et al. [8]. Comparison was done based on the reported accuracy and speed of the method, the level of human involvement in the process of calibration, and special equipment or limitations required by the method. Overall, our method is superior in all categories, as shown in Table 1. The acquisition and processing times of Hu et al. [8] were estimated based on the fact that they utilized the same algorithm as Jaffe-Dax et al. [10], adding an extra step of manually selecting the optodes locations on the 3D reconstruction. For the inter-method error of Hu et al. [8] we calculated the STD from their provided table. Hu et al. [8] use the SfM algorithm which relies on SIFT to perform feature matching, yet it does not perform

well with visually uniform patches [22] and is, therefore, highly susceptible to the cap model. We assume the authors had to hand-tailor the visual appearance of the cap to a satisfactory degree, an action that requires some computer-vision expertise. Moreover, this would not be robust for other cap types or models. To overcome this challenge, Jaffe-Dax et al. [10] use visual Perlin noise patterns printed on the cap allowing the SfM algorithm to achieve good reconstruction. However, adjusting the cap properties still does not address other appearance based noise sources (e.g. occlusions, lighting conditions and quality of video). Also this method forces stickers to be placed on the face of the subject, which is very distracting for the developmental study group. As opposed to these methods, we avoid the need to find good matching features all-together, while still providing robustness to visual appearance using convolutional neural networks. Thus, our method is useful for any cap in numerous different environments, as long as green stickers appear in the video and without the need to place them on the face of the subject. Retraining also makes it possible to work with other colors, and yield even better results if needed, by annotating as little as 300 images manually (the training procedure is described in the supplementary material). Finally, in contrast to the other solutions, our method's speed allows fast re-calibration on the spot when the taken video fails to meet the essential requirements (e.g. wrong camera path, the subject moved too much, occlusions, etc.). For adults, the cap can even be adjusted in real-time for better positioning on the scalp, using the calibration results rather than using them as a post-processing step (for developmental population, once the cap is placed it is rather challenging to re-position it without distracting the subject).

4 Discussion

We proposed a real-time, automatic, photogrammetry-based method to calibrate positions of optodes on the scalp. When positioning the cap on a subject's head, even minute shifts in its orientation relative to the scalp might translate to large errors in the cortical regions being recorded. Such errors propagate into the aggregation of signals between subjects as noise, and harm the process of conclusion drawing regarding cortical activity.

Our approach uses a pipeline consisting of state-of-the-art neural networks to counter these errors by calibrating the position of the cap in an accurate, timely, and convenient way. We designed this method to overcome many of the disadvantages of traditional techniques, and have shown it can be applied to developmental populations. Our tool opens new possibilities in the field of cognitive neuroscience by allowing out-of-lab calibration, and thus experiments to be performed with very little constraints. This is due to the nature of the data-driven convolutional neural-networks we use and their robustness to many scenarios. In addition, the method provides automatic and real-time calibration, which means the user can get immediate feedback regarding the quality of the calibration and easily perform it again if required.

We presented evidence that support the proposed methods reliability and validity using the widely accepted field's standard - the 3D digitizer. We also showed it is superior to previous photogrammetry based approaches.

Our approach has some limitations: firstly, we used a rigid-body model as a proxy for the cap, which can actually deform and stretch on the scalp in some cases. To address this, we described an offline step which takes custom template models as input and fine-tunes STORM-Net to accommodate for many types of caps. This minimizes cap deformations during inference, keeping the rigid-body model assumptions valid. On top of this, most fNIRS caps enforce a fixed distance between all optodes which further constrains the type of possible deformations it can undergo. Thus, for these types of caps we expect even better performance than we measured. Secondly, using neural networks as a calibration tool potentially entails unclear behavior, since we can never be sure how the decision for any particular calibration was made, and how well it performs on a particular instance of the input. This is a well known and common problem to all neural networks. To address this, we split the task at hand (i.e. from raw video to calibration parameters) into smaller discrete blocks that offer better explainability. This way, if the calibration process fails for some reason, it is much easier to trace back what went wrong (and possibly perform manual corrections using the GUI) in this kind of design rather than using a black-box end-to-end neural network.

To summarize, our easy-to-utilize calibration approach strengthens the major advantage of fNIRS as being the optimal neuroimaging method to measure the cotrical response in naturalistic settings and

developmental populations, by providing a real-time, automatic, and very simple to use method to calibrate the cap to the subject's scalp.

Acknowledgments and Disclosure of Funding

All authors declare that they have no conflicts of interests.

We would like to thank Dor Edry for his python implementation of the MNI transformation as described in [18].

This research was supported by the Israel Science Foundation grant 2434/19.

References

[1] Christopher M. Aasted, Meryem A. Yücel, Robert J. Cooper, Jay Dubb, Daisuke Tsuzuki, Lino Becerra, Mike P. Petkov, David Borsook, Ippeita Dan, and David A. Boas. Anatomical guidance for functional near-infrared spectroscopy: AtlasViewer tutorial. Neurophotonics, 2(2): 1 – 16, 2015. doi: 10.1117/1.NPh.2.2.020801. URL https://doi.org/10.1117/1.NPh.2.2.020801.

[2] G. Bradski. The OpenCV Library. Dr. Dobb's Journal of Software Tools, 2000.

[3] Tommy Clausner, Sarang S. Dalal, and Maité Crespo-García. Photogrammetry-based head digitization for rapid and accurate localization of EEG electrodes and MEG fiducial markers using a single digital SLR camera. Frontiers in Neuroscience, 11, May 2017. doi: 10.3389/fnins.2017.00264. URL https://doi.org/10.3389/fnins.2017.00264.

[4] Lauren L. Emberson, John E. Richards, and Richard N. Aslin. Top-down modulation in the infant brain: Learning-induced expectations rapidly affect the sensory cortex at 6 months. Proceedings of the National Academy of Sciences, 112(31):9585–9590, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1510343112. URL https://www.pnas.org/content/112/31/9585.

[5] FASTRAK. Fastrak digitizer. https://polhemus.com/_assets/img/FASTRAK_Digitizer_Brochure.pdf. Polhemus, Colchester, Vermont, USA.

[6] Xiaoxue Fu and John E Richards. Age-related Changes in Diffuse Optical Tomography Sensitivity to the Cortex in Infancy. bioRxiv, 2020. doi: 10.1101/2020.08.22.262477. URL https://www.biorxiv.org/content/early/2020/08/23/2020.08.22.262477.

[7] Simon Homölle and Robert Oostenveld. Using a structured-light 3D scanner to improve EEG source modeling with more accurate electrode positions. Journal of Neuroscience Methods, 326(February):108378, 2019. ISSN 1872678X. doi: 10.1016/j.jneumeth.2019.108378. URL https://doi.org/10.1016/j.jneumeth.2019.108378.

[8] Xiao-Su Hu, Neelima Wagley, Akemi Tsutsumi Rioboo, Alexandre F. DaSilva, and Ioulia Kovelman. Photogrammetry-based stereoscopic optode registration method for functional near-infrared spectroscopy. Journal of Biomedical Optics, 25(9):1 – 13, 2020. doi: 10.1117/1.JBO.25.9.095001. URL https://doi.org/10.1117/1.JBO.25.9.095001.

[9] Farzin Irani, Steven M Platek, Scott Bunce, Anthony C Ruocco, and Douglas Chute. Functional Near Infrared Spectroscopy (fNIRS): An Emerging Neuroimaging Technology with Important Applications for the Study of Brain Disorders. The Clinical Neuropsychologist, 21(1):9–37, 2007. doi: 10.1080/13854040600910018. URL https://doi.org/10.1080/13854040600910018.

[10] Sagi Jaffe-Dax, Amit H. Bermano, Yotam Erel, and Lauren L. Emberson. Video-based motion-resilient reconstruction of three-dimensional position for functional near-infrared spectroscopy and electroencephalography head mounted probes. Neurophotonics, 7(3):1 – 14, 2020. doi: 10.1117/1.NPh.7.3.035001. URL https://doi.org/10.1117/1.NPh.7.3.035001.

[11] C. Kabdebon, F. Leroy, H. Simmonet, M. Perrot, J. Dubois, and G. Dehaene-Lambertz. Anatomical correlations of the international 10-20 sensor placement system in infants. NeuroImage, 99:342–356, 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2014.05.046. URL http://dx.doi.org/10.1016/j.neuroimage.2014.05.046.

[12] Davis E. King. Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, 10:1755–1758, 2009.

[13] Sarah Lloyd-Fox, John E. Richards, Anna Blasi, Declan G. M. Murphy, Clare E. Elwell, and Mark H. Johnson. Coregistering functional near-infrared spectroscopy with underlying cortical areas in infants. Neurophotonics, 1(2):1 – 16, 2014. doi: 10.1117/1.NPh.1.2.025006. URL https://doi.org/10.1117/1.NPh.1.2.025006.

[14] J. L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, volume 3, pages 314–317 vol.3, 2000.

[15] Paola Pinti, Ilias Tachtsidis, Antonia Hamilton, Joy Hirsch, Clarisse Aichelburg, Sam Gilbert, and Paul W. Burgess. The present and future use of functional near-infrared spectroscopy (fnirs) for cognitive neuroscience. Annals of the New York Academy of Sciences, 1464(1):5–29, 2020. doi: https://doi.org/10.1111/nyas.13948. URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.13948.

[16] Valentina Quaresima, Silvia Bisconti, and Marco Ferrari. A brief review on the use of functional near-infrared spectroscopy (fNIRS) for language imaging studies in human newborns and adults. Brain and Language, 121(2):79–89, 2012. ISSN 0093-934X. doi: https://doi.org/10.1016/j.bandl.2011.03.009. URL http://www.sciencedirect.com/science/article/pii/S0093934X11000587.

[17] Gerald Russell, K Jeffrey Eriksen, Pieter Poolman, Phan Luu, and Don Tucker. Geodesic photogrammetry for localizing sensor positions in dense-array eeg. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology, 116:1130–40, 06 2005. doi: 10.1016/j.clinph.2004.12.022.

[18] Archana K. Singh, Masako Okamoto, Haruka Dan, Valer Jurcak, and Ippeita Dan. Spatial registration of multichannel multi-subject fnirs data to mni space without mri. NeuroImage, 27 (4):842 – 851, 2005. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2005.05.019. URL http://www.sciencedirect.com/science/article/pii/S1053811905003319.

[19] Olga Sorkine-Hornung and Michael Rabinovich. Least-squares rigid motion using svd. Technical report, ETH Zurich, Department of Computer Science, 01 2017.

[20] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[21] Xuetong Zhai, Hendrik Santosa, and Theodore J. Huppert. Using anatomically defined regions-of-interest to adjust for head-size and probe alignment in functional near-infrared spectroscopy. Neurophotonics, 7(3):1 – 24, 2020. doi: 10.1117/1.NPh.7.3.035008. URL https://doi.org/10.1117/1.NPh.7.3.035008.

[22] Liang Zheng, Yi Yang, and Qi Tian. SIFT meets CNN: A decade survey of instance retrieval. CoRR, abs/1608.01807, 2016. URL http://arxiv.org/abs/1608.01807.