Title: A Camera-based Contactless Approach to the Detection of Stroke Recovery Actions

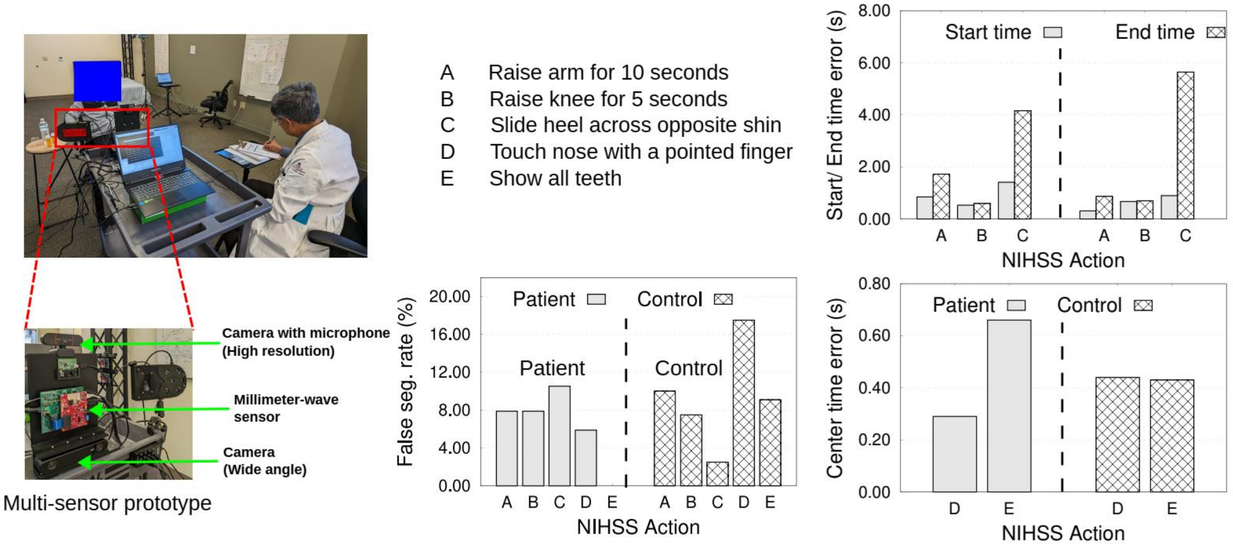
Authors: Moh Sabbir Saadat, Sankalp Jajee, Sanjib Sur, Souvik Sen

Objectives: Post-stroke recovery is lengthy, stressful, painful, and costly. The standard recovery tracking procedures (NIH Stroke Scale, Fugl-Meyer Test, etc.) involve in-person appointments with doctors or trained therapists. This adds to the stress on the recovering survivors who are often wheelchair-bound or are suffering from some extent of disability. The procedures, in general, are composed of a series of tests to assess motor functions, facial palsy, visual field of view, level of consciousness, etc. Fortunately, recent advances in AI/ML enable much finer estimation of facial and postural landmarks from sensing systems such as a camera. To this end, a camera-based detection of stroke recovery actions can enable automated assessment of stroke severity from commodity devices (e.g. a laptop), removing the need for in-person appointments. Moreover, the associated microphone can pick up voice cues whereas additional sensors such as wireless networking signals can bolster the camera outputs for even more fine-grained detection and scoring.

Methods: We developed a multi-sensor prototype with several synchronized sensors, including an RGB camera with a microphone. The goal is to capture a real NIH Stroke Scale (NIHSS) procedure between a doctor and a subject with the sensors so that the sensor inputs can be used to detect the beginning and end of the various NIHSS actions. Our approach is to estimate the coordinates of postural and facial landmarks on RGB image frames and then, formulate a suitable feature vector for each action in the NIHSS procedure that allows the temporal segmentation of the action. However, it is challenging to search across the entire length of the procedure for a specific action segment since the actions often have similarities with normal activities of a sitting person, *e.g. the subject pointing at a picture is similar to raising his/her arm, scratching the face or nose is similar to touching nose, etc.* To overcome this ambiguity, we utilize the microphone inputs and the recent advances in Automated Speech Recognition (ASR) to localize, in time, the command that corresponds to the specific action. This gives us an initial coarse filter for the action on the temporal domain.

Results: Ten stroke survivors (“patient”, mean age: 62.4 years, standard deviation: 11.8 years) and ten healthy individuals (“control”, mean age: 31.6 years, standard deviation: 17.7 years) participated in the study, and a doctor performed two cycles of NIHSS procedure on each participant. Using our microphone-camera segmentation approach, we obtained **false segmentation rates**, *i.e. the time window of the action was not correctly identified*, of 7.89%, 7.89%, 10.53%, 5.88%, and 0.00% for the patients across the NIHSS actions A to E (description in figure below). The corresponding false segmentation rates for the control subjects were 10.00%, 7.50%, 2.50%, 17.50%, and 9.09%. The **(start time errors, end time errors)** across actions A to C are (0.85 s, 1.72 s), (0.53 s, 0.59 s), and (1.41 s, 4.16 s) for the patients, and (0.31 s, 0.87 s), (0.67 s, 0.70 s), and (0.90 s, 5.64 s) for the controls. The **center time errors** for the actions D and E are 0.29 s and 0.66 s for patients whereas, 0.44 s and 0.43 s for the controls. (Actions D and E span much shorter, momentary durations; thus, error in center time is more appropriate).

Conclusion: In this study, we achieved the first step towards automating an NIHSS procedure using a microphone-camera pair. However, a purely vision-based system often fails to pick up postural and facial signatures reliably as shown by the false segmentation rates. Firstly, AI processing on camera inputs is affected by lighting conditions – *dimmer environments, sunshine glare, etc.* Moreover, certain combinations of skin and clothing colors and the background can create ambiguities in the model. To circumvent these effects, our next approach is to fuse wireless signals from a 5G networking system with camera images. Besides the microphone and camera, our prototype includes a millimeter-wave signal source: millimeter-wave is a core technology for upcoming 5G networks, and it has shown increasing potential for picking up useful signatures from human activities. Wireless signals such as millimeter-wave are unaffected by lighting conditions (no drop in performance even in complete darkness) and they are also blind to color combinations.



Multi-sensor prototype