Semiautomatic marker tracking of tongue positions captured by videofluoroscopy during primate feeding

Matthew D. Best[1] *Student Member, IEEE*, Yuki Nakamura[2,3], Nicoletta A. Kijak[2], Mitchell J. Allen[4], Teresa E. Lever[4], Nicholas G. Hatsopoulos[1,2], Callum F. Ross[2], and Kazutaka Takahashi[2] *Senior Member, IEEE*

***Abstract*— Videofluoroscopy (VF) is one of the most commonly used tools to assess oropharyngeal dysphagia as well as to visualize musculoskeletal structures of humans and animals engaged in various behaviors, including feeding. Despite its importance in clinical and scientific use, processing VF data has historically been extremely tedious because it is performed using manual frame-by-frame methods. With recent technological advances, the frame rate for scientific use has been increasing along with the use of high speed data capture systems. In the current study, we used non-human primates as a model animal to study human feeding behaviors to capture tongue movement based on markers implanted into the tongue. Here, we introduce a semi-automatic marker tracking algorithm that yields high tracking accuracy ($> 90\%$) and dramatic speed improvements (faster than real time labeling). Furthermore, we quantify the sources of tracking errors and the tracking performance as a function of marker speeds. Our results indicate that there is more room for methodological improvements both in detection and prediction of marker positions. Moreover, correspondingly faster frame rates will be required to capture faster kinematic behaviors such as those of mice, which are extensively used to study both control and pathological conditions.**

I. Introduction

Marker-based motion capture technology has been widely and successfully used to obtain kinematic data for various types of behaviors involving the lower and upper limbs/extremities, where reflective markers can be relatively easily placed on subjects of various sizes, including humans, horses, and rodents. However, imaging methods such as ultrasound [1], computed tomography CT [2] and videofluoroscopy (VF) are essential to study feeding and swallowing behaviors in which tissue surfaces are not directly visible [3], [4], [5], [6]. VF is commonly used to assess swallowing impairment (dysphagia) in major neurological diseases such as stroke [7], Parkinsons' disease [8], and amyotrophic lateral sclerosis (ALS) [9], just to name a few. Dysphagia often not only compromises the quality of life of patients but also causes lethal consequences due to malnutrition and aspiration pneumonia [10], [11]. Non-human primate models exhibit similar kinematic behaviors to those by human subjects, particularly tongue and swallowing kinematics, and it is feasible to place markers in the monkeys.

A major bottleneck of VF image analysis, especially compared to the surface marker based motion capture approach, is a lack of established methods for feature tracking. In the current study, we recorded VF data during manual feeding of a non-human primate with markers implanted in the tongue to characterize tongue kinematics as a part of the whole feeding sequence analysis, including swallowing. Then we developed a method of semiautomatic tracking of markers using the speeded up robust features (SURF) algorithm [12] coupled with a Kalman filter to make predictions about the spatial location of tongue markers. Furthermore, we quantified the sources of errors in tracking as well as tracking performance against marker speeds. Our method yielded significant improvement in time to track markers compared to manual labeling by well-trained researchers. Additionally, our speed-tracking performance analysis suggested that a faster capture system and further algorithmic developments are necessary to reliably capture the high speed kinematics of feeding and swallowing in monkeys and other animal models of swallowing, particularly mice which are better suited for studying particular diseases.

II. Methods

A. Experimenal Procedure and Behavioral task

All of the surgical and experimental procedures were approved by the University of Chicago Animal Care and Use Committee and conformed to the principles outlined in the Guide for the Care and Use of Laboratory Animals (NIH publication no 86-23, revised 1985). One female rhesus macaque (Macaca mulatta) was trained to be seated by using a pole and collar technique. The neck collar was removed during training and data recording sessions. The head was restrained with a halo coupled to the cranium through chronically implanted head-posts. At least one month prior to data recording, three 0.5 mm diameter tantalum balls (RSA Biomedical, Sweden) were implanted by hypodermic needle into the anterior, middle, and posterior regions of the tongue at midline, under isoflurane anesthesia. Post-surgery, the monkey was trained to perform a manual feeding task and eat various types of foods in preparation for VF testing.

*This work was supported by NIH grant R01 NS045853, R01 DE023816, R21 NS084870, and Mizzou advantage funding.

[1]M.D.B. and N.G.H. are with the Committee on Computational Neuroscience, University of Chicago, Chicago, IL 60637, USA (e-mail: {mattbest, nicho}@uchicago.edu)

[2]N.A.K., Y.N, N.G.H. C.F.R. and K.T. are with the Department of Organismal Biology and Anatomy, University of Chicago, IL 60637, USA (e-mail: {nkijak, rossc, kazutaka}@uchicago.edu)

[3]Y.N.is with the Graduate School of Medicine and Dental Science, Niigata University, Niigata, Niigata, Japan

[4]M.J.A. and T.E.L. are with the Department of Otolaryngology-Head and Neck Surgery, University of Missouri School of Medicine, Columbia, MO 65212, USA (e-mail: {AllenMJ,levert}@health.missouri.edu).

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

B. Videofluoroscopy and Post-processing

While the monkey self-fed various foods, two-dimensional lateral view VF recordings of jaw and tongue movements were acquired at a frame rate of 100 Hz using an OEC 9600 C-arm fluoroscope retrofitted with a Redlake Motion Pro 500 video camera (Redlake MASD LLC, San Diego, CA) [6]. All data used in subsequent analyses were manually labeled by trained researchers using Pro Analyst (Xcitex, Woburn, MA) to provide a gold standard for comparison with our novel automated methods.

C. Marker Tracking Algorithm

We developed a semi-automatic algorithm to detect the position of the three tongue markers. All computational analysis and tracking was performed within the Matlab (The Mathworks, Natick, MA) computing environment using the computer vision toolbox.

1) Feature Detection: We used the "speeded up robust features" (SURF) algorithm [12] to identify visual features of the fluoroscopy data that were putatively tongue markers. This algorithm identifies scale invariant blob-like features of images by analyzing images at multiple spatial scales. Because the tongue markers were relatively small, we used only fine scale spaces. Particularly, the four spatial filters used by the SURF algorithm were of size 9×9, 15×15, 21×21, and 27×27 pixels where the 9×9 and 27×27 pixel filter captured the finest and coarsest spatial information, respectively.

On every frame, the SURF algorithm identified $N = 25$ interest points (Fig. 1B) that contained visual features consistent with those of the tongue markers.

2) One-step Ahead Prediction: A Kalman filter [13] was used to predict the spatial location of each tongue marker on a frame-by-frame basis. At time t, the position of tongue marker, m was predicted from the previous frames using the following equations:

$$x_m(t) = Ax_m(t-1) + w_m(t)$$
$$z_m(t) = Hx_m(t) + v_m(t)$$

where $x_m(t)$ is the state of tongue marker m at time t; A is the state transition matrix; $w_m(t)$ is process noise; H is the measurement model; $v_m(t)$ is measurement noise at time t, and $z_m(t)$ is the estimated position of tongue marker m at time t (Fig. 1C).P

3) Matching Detected Features to Predictions: Ascertaining which, if any, of the N interest points correspond to the P predicted locations of the tongue markers can be formulated as a combinatorial optimization problem. This class of problem is commonly referred to as the assignment problem, and is solved in polynomial time [14]. In assignment problems, there is a cost, c_{np}, to assign interest point $n \in N$, to prediction $p \in P$. Here, we used the Euclidean distance between n and p as the cost to assign interest point n to prediction p. We computed the pairwise distance between all pairs of interest points and predicted marker locations to generate a cost matrix, $\mathbf{C}$. If the distance between all of the interest points and the estimated tongue position exceeded a threshold (here, 15 pixels), then no interest point was assigned to that tongue marker. We used an optimized algorithm to solve this assignment problem with $O(n^3)$ worst case and $O(n)$ average case performance, respectively (Fig.

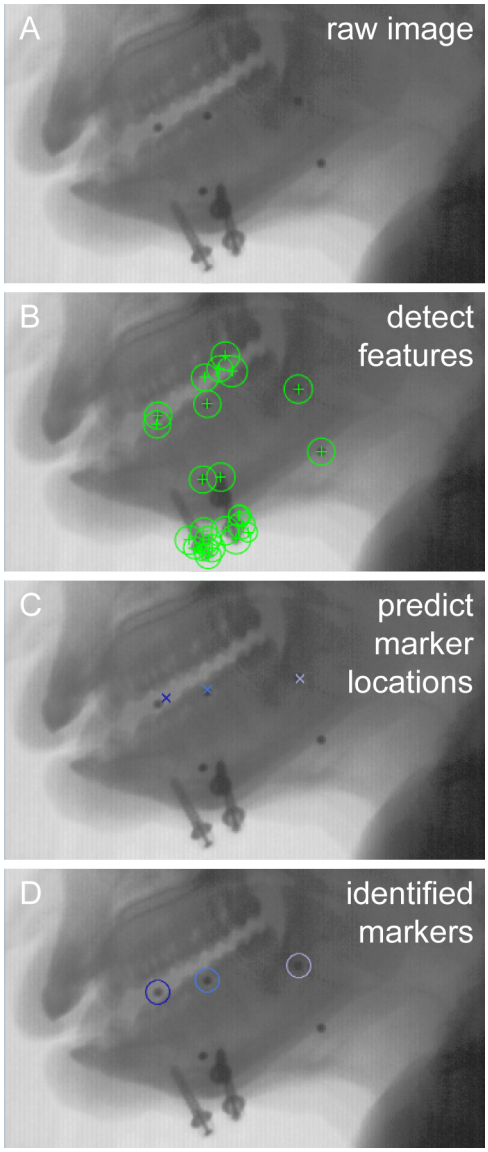

Fig. 1. **Overview of tongue tracking algorithm A.** An exemplar frame of fluoroscopy data. Note the three opaque markers implanted in the tongue in anterior, middle, and posterior locations. **B.** The SURF algorithm was applied to the fluoroscopy image to identify interest points that putatively correspond to tongue markers. The 25 most salient visual features identified by the algorithm are shown as green circles with crosses in their center. Many salient features identified were of jaw screws and other markers that were not implanted on the tongue. Each of the three tongue markers were detected. **C.** A Kalman filter was used to make predictions about the spatial location of the tongue markers based on their location, velocity, and acceleration in previous frames. The predicted location of each tongue marker is indicated by a blue symbol ($\times$). A dark-to-light color gradient indicates the anterior-to-posterior axis of the tongue. **D.** A matching algorithm was used to assign interest points (**B**) to the predicted marker locations (**C**). The interest point that was identified as a tongue marker is shown by a blue circle whose color corresponds to the predictions in **C**. Each tongue marker was successfully identified by the algorithm.

1D) [15].

4) Advancing to the Next Frame: After algorithmically predicting the location of tongue markers, a trained research analyst verified the correctness of the prediction. If one of the tongue markers was not assigned an interest point, the researcher manually labeled the position of the marker. Similarly, if the algorithm incorrectly identified one of the tongue markers, it was manually corrected. After manual verification/correction, the Kalman filter was updated to predict the next frame.

III. Results

We applied our tongue tracking algorithm to several primate feeding sequences. On a limited subset of trials, our algorithm failed to correctly predict the position of one or more tongue markers. These frames may have been incorrectly labeled because the feature detector did not find the tongue marker (Fig. 2A) or because the predicted location of the tongue marker greatly deviated from its actual position, resulting in a different interest point being incorrectly assigned as a tongue marker (Fig. 2B). We will henceforth refer to these two types of errors as detection and prediction errors, respectively.

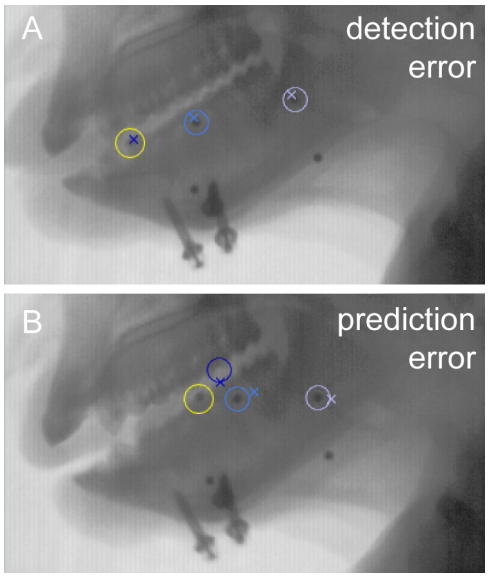

Fig. 2. **Two types of tracking error A.** A representative frame containing a detection error of the anterior tongue marker. The actual location of the tongue marker based on labeling by a trained researcher is indicated by the yellow circle (all other conventions same as Fig. 1). Note that the Kalman filter predicted the marker would be very close to its actual location. **B.** A representative frame showing a prediction error of the anterior tongue marker. Here, the predicted location of the tongue marker (dark blue ×) was far from its actual location (yellow circle). Instead, a different interest point was incorrectly identified as the tongue marker.

We found that an overwhelming majority of frames were correctly identified by our algorithm. Of the 1896 fluoroscopy frames that were analyzed, we correctly identified the position of the anterior, middle, and posterior tongue markers on 1734 (91.5%), 1827 (96.4%), and 1877 (99.0%) frames, respectively (Fig. 3). Detection errors occurred on 89 (4.7%), 34 (1.8%), and 18 (1%) frames, while prediction errors were present on 73 (3.9%), 35 (1.9%), and 1 (0.05%) frames for the three tongue markers, respectively.

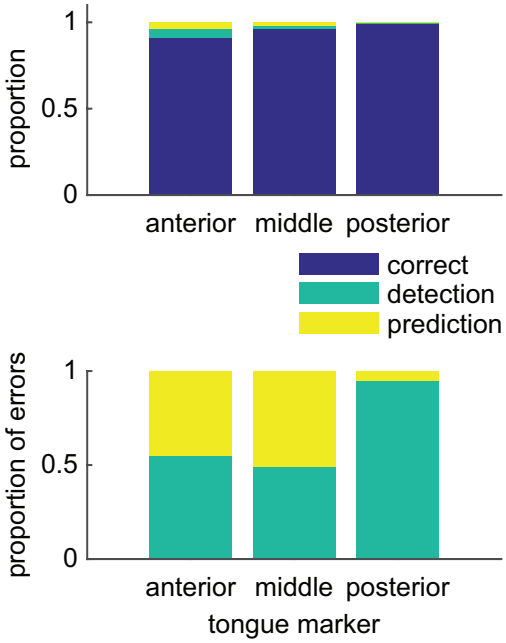

Fig. 3. **Tracking outcomes** We correctly identified the location of the tongue marker on over 90% of frames (top). On frames that contained an error, we quantified the proportion that were due to detection errors and prediction errors. For the anterior and middle tongue markers, detection and prediction errors occurred in roughly equal proportion. For the posterior tongue marker, detection errors comprised nearly all of the total errors.

We calculated tracking performance as a function of tongue speed and found that in each instance, our ability to track tongue markers worsened as the tongue moved faster (Fig. 4).

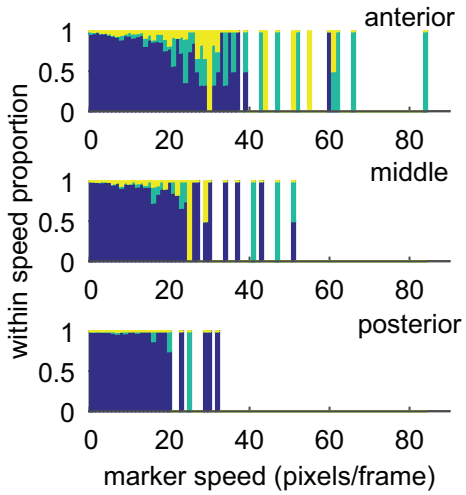

Fig. 4. **Tracking performance as a function of tongue speed** We found that both detection and prediction errors were more likely to occur as the tongue moved faster. Color scheme is the same as Fig. 3.

We compared the amount of time trained researchers spent manually labeling the tongue tracking data with the amount of time it took a trained researcher to verify and correct

mislabeled algorithmic data. We found that the labeling rate, expressed as frames labeled per unit time, was nearly 500% faster for algorithmically labeled data.

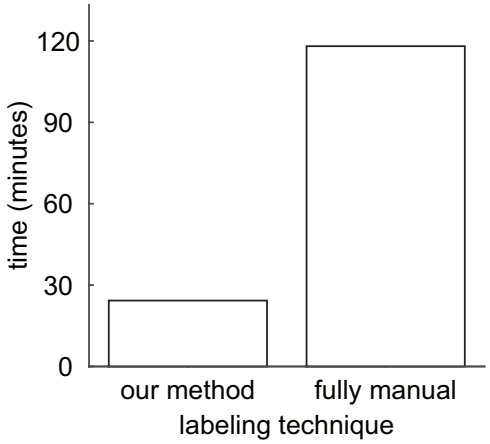

Fig. 5. **Improvement in tracking speed** We compared the amount of time it took a human to label the tongue tracking data using a manual procedure with our semi-automatic tracking procedure. We found that algorithmic tracking of tongue position was nearly 500% faster than manual labeling.

IV. Discussion

Our method yielded a satisfactory tracking performance of markers in much less time than human intervention. Furthermore, the error classification analysis suggests that both the detection and prediction components can be further improved. In the current study, we did not manipulate the image quality of the VF data. Thus, preprocessing of images such as contrast adjustments should improve the detection performance. However, in the present study, there was an inherent blurring of frames due to the high speed kinematics of feeding and swallowing relative to the slower camera frame rate. The anterior marker showed a much wider range of speed, and the faster the speed of the marker, the worse the tracking performance became. This kinematic speed - frame speed relation will be a critical issue to overcome when studying mice, a more common animal model in feeding and swallowing research. The wide availability of both control and diseased models of mice, in conjunction with established experimental protocols to obtain VF data [16], make the development of VF analysis tools for mice an essential step in the progression of neurodegenerative disease research. Behaviorally, tongue movements of mice are much faster than the tongue kinematics of the monkey included in our study, with mice reaching lick rates of 7-8Hz or higher [16], [17]. In order to capture relevant kinematics from mice, it will be essential to develop a much faster VF capture system than is currently available, and based on our analysis, the frame rate needs to be at least a few times faster than the 100 frames/sec camera used in our study. Furthermore, although most VF studies are limited to two dimensions, most of the movements analyzed with VF are inherently three dimensional. New devices have been developed to capture three dimensional morphological data during mammalian feeding [18]. Our future work will incorporate constraints posed by two camera images in three dimensional imaging and geometric constraints of marker locations in relation to rigid body elements in animal models. In order to make clinical translation of our work to connect our animal models and human subjects, we are in the process of detecting dynamic features without markers such that soft tissue kinematics and food/liquid bolus flow can be detected and predicted with high accuracy and compared against data with markers.

Acknowledgment

This work was completed in part with resources provided by the University of Chicago Research Computing Center.

References

[1] Wylie J. Dodds, "The physiology of swallowing," *Dysphagia*, vol. 3, no. 4, pp. 171–178, 1989.

[2] C. M. Orr, E. L. Leventhal, S. F. Chivers, M. W. Marzke, S. W. Wolfe, and J. J. Crisco, "Studying primate carpal kinematics in three dimensions using a computed-tomography-based markerless registration method," *Anatomical Record*, vol. 293, no. 4, pp. 692–701, 2010.

[3] S. E. Langmore, K. Schatz, and N. Olson, "Endoscopic and videofluoroscopic evaluations of swallowing and aspiration.," *The Annals of otology, rhinology, and laryngology*, vol. 100, no. 8, pp. 678–681, Aug. 1991.

[4] M Mahesh, "Fluoroscopy: patient radiation exposure issues.," *Radiographics : a review publication of the Radiological Society of North America, Inc*, vol. 21, no. 4, pp. 1033–1045, 2001.

[5] M. G. Rugiu, "Role of videofluoroscopy in evaluation of neurologic dysphagia.," *Acta otorhinolaryngologica Italica*, vol. 27, no. 6, pp. 306–316, 2007.

[6] C. F. Ross, A. L. Baden, J. Georgi, A. Herrel, K. A. Metzger, D. A. Reed, V. Schaerlaeken, and M. S. Wolff, "Chewing variation in lepidosaurs and primates," *The Journal of Experimental Biology*, vol. 213, no. 4, pp. 572–584, 2010.

[7] R. Shaker and J. E. Geenen, "Management of Dysphagia in Stroke Patients," May 2011.

[8] M. J. Waxman, D. Durfee, M. Moore, R. A. Morantz, and W. Koller, "Nutritional aspects and swallowing function of patients with Parkinson's disease.," *Nutrition in clinical practice*, vol. 5, no. 5, pp. 196–199, 1990.

[9] J. Robbins, "Swallowing in ALS and motor neuron disorders.," *Neurologic clinics*, vol. 5, no. 2, pp. 213–229, May 1987.

[10] J. A. Y. Cichero and B. E. Murdoch, Eds., *Dysphagia: Foundation, Theory and Practice*, Wiley, 1st edition, 2006.

[11] R. Leonard and K. Kendall, *Dysphagia Assessment and Treatment Planning: A Team Approach*, Plural Publishing Inc, 3rd edition, 2013.

[12] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[13] R. E. Kalman, "A New Approach to Linear Filtering & Prediction Problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[14] J. Munkres, "Algorithms for the Assignment and Transportation Problems," *Journal of the SIAM*, vol. 5, no. 1, pp. 32–38, 1957.

[15] M. L. Miller, "Optimizing Murty s Ranked Assignment Method," *Ieee Transactions On Aerospace And Electronic Systems*, vol. 33, no. 3, pp. 851–862, 1997.

[16] T. E. Lever, R. T. Brooks, L.A.. Thombs, L. L. Littrell, R.A. Harris, M.J. Allen, M. D. Kadosh, and K. L. Robbins, "Videofluoroscopic Validation of a Translational Murine Model of Presbyphagia," *Dysphagia*, , no. 1, pp. 6–10, 2015.

[17] J.A.W.M. Weijnen, "Licking behavior in the rat," *Neuroscience and Biobehavioral Reviews*, vol. 22, no. 6, pp. 751–760, 1998.

[18] E. L. Brainerd, D. B. Baier, S. M. Gatesy, T. L. Hedrick, K. A. Metzger, S. L. Gilbert, and J. J. Crisco, "X-ray reconstruction of moving morphology (XROMM): precision, accuracy and applications in comparative biomechanics research.," *Journal of experimental zoology. Part A, Ecological genetics and physiology*, vol. 313, no. 5, pp. 262–279, 2010.