# Robotic cooking through pose extraction from human natural cooking using OpenPose

Dylan Danno[1], Simon Hauser[1], and Fumiya Iida[1]

[1] Bio-Inspired Robotics Laboratory (BIRL), Department of Engineering
Cambridge University, United Kingdom, fi224@cam.ac.uk

**Abstract.** Robotic cooking is a difficult task, as the translation from raw recipe instructions to robotic actions involving precise motions and tool-handling is challenging. This paper introduces automated recipe technique recognition based on recording the pose trajectories of human demonstrators carrying out a pancake recipe, as an intuitive method for untrained demonstrators to show a robot how to carry out their recipe variant. A Kinect 2 sensor and the OpenPose neural network are used to extract key timings from the demonstrations, which are replicated when the recipe is carried out by a UR5 arm. Comparing several human-cooked pancakes with their robot-replicated counterparts, the robot's pancake quality scores were found to be only slightly inferior to the human ones, suggesting that the key parameters selected did encompass the most important variations in cooking technique. Furthermore, we discuss preliminary results in inferring the relationship between the cooking parameters and the quality scores.

**Keywords:** domestic robots, robotic cooking, tracking, OpenPose

## 1 Introduction

Robotic automation of culinary tasks is an ongoing challenge, with several prototype robot chefs under development. These include cookie baking robots used to investigate sensing and manipulation in the kitchen environment [1], pancake making robots that operate off of written instructions [2], and robotic kitchens that autonomously produce restaurant quality food [3]. The prospect of cooking robots in the average home is moving closer to reality. Given the amount of variation seen in human cooking, domestic end users will likely want their robot chefs to carry out custom variants of recipes, which is problematic given that the average domestic cook does not have extensive programming or robotics experience. This problem of how to transfer important cooking information from the end user to the cooking robot is challenging, and also raises the question of what recipe information is in fact important.

One intuitive way of transferring the desired recipe information is teaching from human demonstrations. The more general prospect of teaching robots by demonstration is the subject of considerable prior work [4, 5, 6, 7, 8]. Prior work also exists on applying this technique to cooking, but has often used only 2D

video footage [9], or collected 3D information in manner not suitable for the domestic environment, such as requiring a battery of cameras and the demonstrator to wear bulky pose tracking equipment [3].

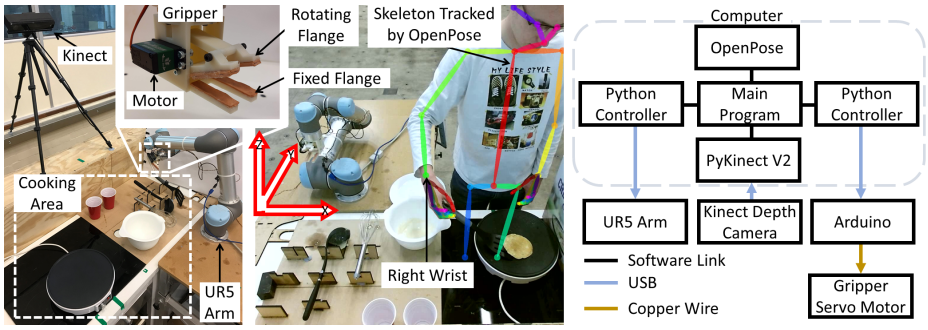


**Fig. 1.** Left: experimental setup with Kinect camera, UR5 arm, and cooking area with the necessary tools. The inset shows details of the used gripper for tool manipulation. Middle: view from the Kinect camera (world coordinate axes superimposed), tracking the human skeleton, specifically the right wrist. Right: schematic of the hardware and software used

This project investigated replicating variants in demonstrators' cooking techniques using a single sensor's 3D recordings of their mostly unconstrained cooking. This approach is suited to the domestic environment, being both low cost and markerless. As a task that can be implemented with our present hardware while still involving sufficiently complex actions, we chose pancake-making as the goal of our robotic chef. At first, the system observes several human cooks performing the recipe, using an RGB-D camera and OpenPose to track their skeletons (see Fig. 1). In general, these cooking trajectories are rich in features, even for a relatively simple recipe consisting of only a few limited recipe instructions: picking up and returning tools, complex tool-handling sequences, and observing idle time smoothly transition between one another. In our preliminary investigation, among all possible such cooking variables, the system analyses the obtained trajectories for only how much *time* each cook spent on the specific subtasks of the recipe. We hypothesise that the final quality and subjective assessment of the cooked pancake highly correlates with subtask timings, while how the actual actions are executed is of lesser importance. This simplifies the complex cooking trajectories to a sequence of actions that can be precoded into robotic hardware and whose execution time can be varied. We developed an algorithm which automatically extracts the timing of pre-defined subtasks from the tracked trajectories. The most important of these times are then replicated by the robot, and the obtained pancakes from both the human and robotic chefs are then evaluated in a blind test. We assume that in order for a human chef to be satisfied with the quality of the food cooked by the robotic chef, it needs to be indistinguishable from the human chef's variant. Hence, our goal in this

initial research is not to optimise the timing or any other parameter per se, but to understand if human testers are able to distinguish a pancake cooked by a human or a robot with the same *timings* for the subtasks. This could provide insight to robotic chefs about which processes need to be matched closely to the style the demonstrator used, and which processes could potentially be solved by only recreating the action in a non-human manner (i.e. in a manner simpler for the robotic hardware). Our approach is not limited to only pancake-cooking and thus could be extended for future research on robotic cooking.

## 2 Materials and Methods

In this section, we detail the processes from human chefs cooking a pancake to the robotic chef performing the same task, as laid out in Fig. 2. The full process involves four steps: recording, trajectory analysis, replication, and evaluation. The following sections explain each step in detail.

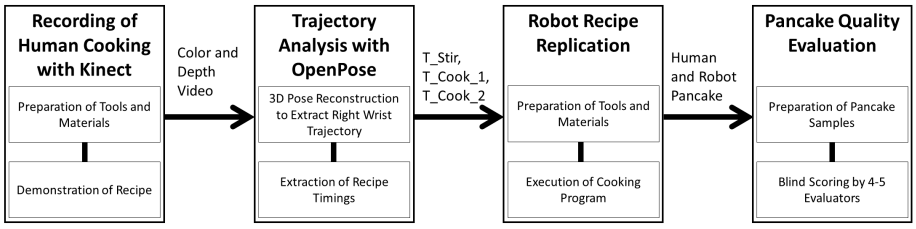


**Fig. 2.** Process Overview. First, a human chef cooking a pancake is recorded. The trajectory is passed to an analysis process which extracts key timings according to a developed algorithm. The robotic chef then replicates the pancake cooking with these timings, upon which both pancakes are evaluated by human testers in a blind test.

### 2.1 Recording of Human Cooking

**Recipe and Setup** For timing purposes, the pancake recipe has been divided into seven stages (see Fig. 3 left). Among them, three stages of the recipe are deemed to be time sensitive, as they intuitively influence the final quality of the pancake: batter stirring (T_Stir), cooking before flipping (T_Cook_1), and cooking after flipping (T_Cook_2). We hypothesise that the timings of these stages are the most important aspects of the recipe, and so that replicating them exactly will result in pancakes of similar quality to that of the human demonstrator.

The setup is shown in Fig. 1 left and middle. The relevant tools - cups, whisk, ladle and spatula - are held in dedicated mounts on the table, alongside a fixed mixing bowl and an electric hotplate. On the one hand, fixed tool positions simplifies the trajectory analysis, as will be explained in the following section.

On the other hand, due to hardcoded pick-up and return tools sequences used in this work for the robotic chef, exact and deterministic tool positions are required for a successful recipe replication. During experiments the hotplate is pre-heated to 180° C and pre-greased with vegetable oil.

**Recording of Demonstrations** Each demonstration involves a human chef cooking a pancake according to the predefined recipe. The only constraint on their cooking is a request to use only their right hand during the recipe to ensure that their motions could be replicated by the UR5 arm and its gripper. Given that the tools involved in this recipe are innately single handed, this single constraint only excludes a small number of behaviours (such as e.g. dispensing the pancake mix and water into the bowl simultaneously). Hence this constraint was considered as practically insignificant as the cooking technique observed will likely be the same as the completely unconstrained method the demonstrator would otherwise use. During the demonstrations, a Kinect V2 sensor is used to record colour and depth images of the demonstrator at rate of 10 Hz.


**Fig. 3.** Robot recipe replication summary. Left: Recipe stages. Middle: Corresponding snapshots of human chef cooking. Right: Corresponding snapshots of robotic chef cooking. Some sequences are hardcoded (such as e.g. grasping a cup), while more complex sequences are recorded through "kineaesthetic teaching", denoted as KT.

### 2.2 Trajectory Analysis with OpenPose

**3D Pose Reconstruction to Extract Right Wrist Trajectory** After a human demonstrator is recorded, the OpenPose [10] neural network is provided with the 2D Kinect colour video recording of the demonstration to determine the position of the demonstrator's joints (see Fig. 1 middle). OpenPose carries this out using a "bottom-up" approach, first identifying body parts, then stitching those together to form a continuous human, which adds robustness to occlusion. The output of the network is the 2D joint positions of the observable skeleton in each frame of the colour video (see Fig. 3 middle). To obtain the full 3D pose of the skeleton, the depth information of the Kinect camera can be used. First, the colour frames must be mapped to the properties of the depth frames, for which the PyKinect2 library (a python port of the Kinect for Windows SDK 2.0 C++ Library) was used. Then, each 2D joint position can be associated with the corresponding depth information at that point, resulting in the 3D trajectory of each joint, and thus to full 3D pose reconstruction. The raw pose trajectory of the demonstrator's skeleton is noisy, and so each joint's trajectory has been smoothed using a Savitzky–Golay filter [11] of width 21 datapoints (i.e. 2.1 s) and a polynomial order of 2.

**Extraction of Recipe Timings** This smoothed trajectory is analysed to determine the time spent on each stage of the recipe. This is carried out via a spatial discrimination method - specifically tracking the x-y position (in global coordinate frame) of the right wrist (Fig. 4 top). The right wrist was initially chosen as the tracked point as its detection is more robust to occlusion than that of other points on the hand. While more sophisticated methods for determining hand position could be imagined (e.g. using the position of the knuckles and wrist to define a hand centre point, etc), these were not explored because sufficient accuracy was found to be obtainable from the wrist measurements (the standard deviation of 10 s of the recorded positions of a stationary, unoccluded wrist is less than 0.2 cm overall).

Figure 4 bottom left shows the smoothed planar trajectory of the right wrist during a demonstration of the pancake recipe by a human chef, with the box regions used for spatial discrimination superimposed. The expected order of hand movements through the boxes is known (as the recipe leaves no scope for alternate orders of instructions, e.g. the pancake cannot be flipped before it is poured in the pan etc.), and so the amount of time the demonstrator spends on each stage of the recipe can be recorded using the time at which their hand passes through each box. The boxes themselves have been determined from the positions of the relevant tool holders, dilated by an appropriate amount to account for that the wrist does not move exactly over the centre of a tool when it is grasped.

The bottom right side of Fig 4 shows the example of the wrist trajectory part from which T_Stir is extracted. After having gone through the whisk box to pick up the whisk, T_Stir starts when the wrist enters the box around the bowl to stir it, and ends when the wrist then subsequently leaves that box again. The cooking times are extracted in a similar fashion. After having gone through

the ladle box to grab the ladle and bowl box to scoop the batter, T_Cook_1 starts when the wrist enters the area around the pan. Given by the recipe, it ends when the pancake is flipped. After having gone through the spatula box to grab the spatula for flipping, this is detected by the wrist exits the area around the pan right after flipping the pancake to return the spatula. At the same time T_Cook_2 starts, which in turn finishes when the wrist exits the pan area for the last time with the spatula to retrieve the pancake. Note that such event-based trajectory segmentation is only possible due to the exact definition of recipe sequences; more complex recipes will likely require more elaborate segmentation processes, however this is outside the scope of the current work.

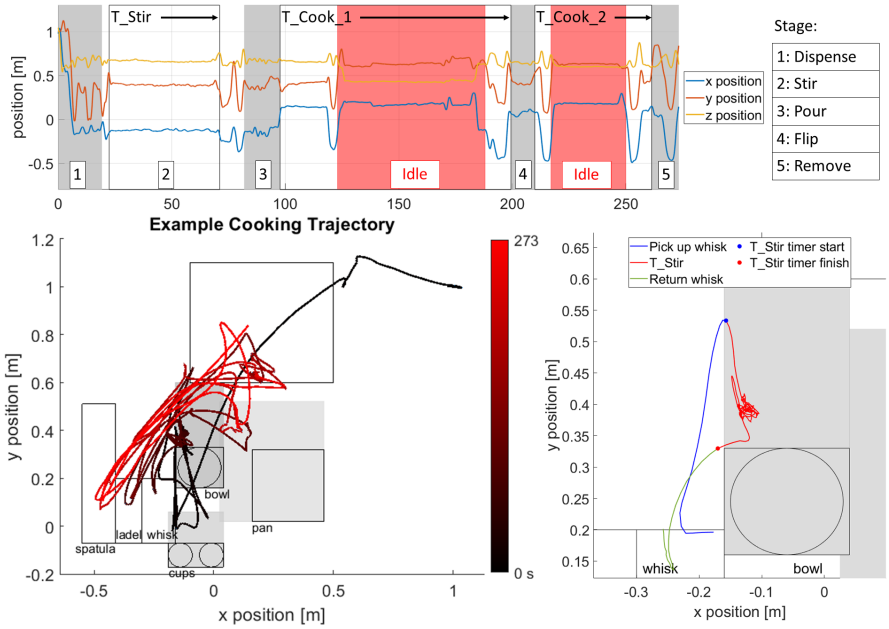


**Fig. 4.** Top: x, y and z trajectories of the demonstrator's right wrist over time. Bottom left: example overall 2D trajectory. Bottom right: example trajectory for extracting T_Stir.

### 2.3 Robotic Chef

**Arm and Gripper** The robot used is a UR5 arm equipped with a simple one degree of freedom gripper controlled by an Arduino board (see Fig. 1 left). The gripper consists of two flat flanges, one of which is attached to a servomotor, which can rotate it to bring it closer to or further from the other fixed hand. The gripper is controlled by commanding the two flanges to move to a certain angle of separation within a given time period. It was fabricated by 3D printing,

and is coated in a polymer material to increase the surface friction to facilitate picking up and holding tools (i.e. the cups, whisk, ladle and spatula). The UR5 arm and gripper are controlled via python software running on a PC.

**Robot Cooking Program** The aim of our strategy is to replicate the important aspects of the demonstration, while carrying out the less important aspects in a way that is easier to automate. The selection of the important aspects of the recipe was done based on intuition as a starting point. The stages of the robot cooking program are shown in Fig. 3 right. The actions involved in the important stages of the recipe have been manually coded with tunable timings, allowing the UR5 to spend the same amount of time on said recipe stages as the human demonstrator. In contrast, the dispense, pour, flip and remove stages are carried out by executing a kinaesthetically taught (KT) trajectory. This is a trajectory taught to the robot by manhandling it through the required motions, with its position recorded at 0.1 s intervals, which is then replayed to reproduce the taught actions. This was done as it was considered that any variations in how these stages were carried out would have minimal effects on the final product. All stages involve the robot gripping and replacing tools in their specific holders, the locations of which have been hard coded. The final robot cooking trajectory thus consists of piecewise hardcoded sequences into which the timing-sensitive actions are slit in.

## 3 Experiments

### 3.1 Human Demonstrations

Five volunteers carried out demonstrations, using the tools visible in Figure 1. Each was given the recipe in case they were unfamiliar with it. This does not bias our results or analysis system, as the recipe cannot be carried out any other way than by the sequence described in Fig. 3.

### 3.2 Evaluation Protocol

After a human demonstration was carried out, their pancake was first allowed to cool, then stored in an airtight container. The UR5 then replicated the recipe with their timings. Once the second pancake had been allowed to cool (to ensure that the two pancakes could not be distinguished by a difference in temperature), the pancakes were cut into portions. Three to four evaluators were then (blindly) given a sample from both the demonstrator's and the robot's pancake, and asked to evaluate it on a scale of 1-10 on the parameters of taste, texture and appearance, with one pancake set to having a reference score of 5-5-5. Unbeknown to the evaluators, the demonstrator's pancake was always used as the reference for consistency between the results. Hence we evaluated quality scores of the robot pancakes relative to the human ones. A video showing the full procedure is available under https://www.youtube.com/channel/UCssbO6gQLdwC2T3ZNkAo1wA.

## 4 Results and Discussion

**Accuracy** We first assessed the accuracy of the proposed timing method. For this, we compared the automatically extracted timings to their ground truth value, manually measured from timestamps in the video recordings of the demonstrations. The differences between the recorded and "true" times were on average 3.1 s (12%) for T_Stir, 1.8 s (2%) for T_Cook_1 and 1.9 s (2%) T_Cook_2 (see Fig. 5 top left), which was deemed acceptable for our purposes. The maximum absolute error seen was 5.6 seconds, 5% of the manually measured T_Cook_2, while the largest percentage error seen was 21%, a 4.1 s error in a 23.9 s T_Stir. The stirring times had larger percentage errors due to their shorter duration, and possessed a overestimation bias, suggesting that the automatic stir timing method could use further refinement.

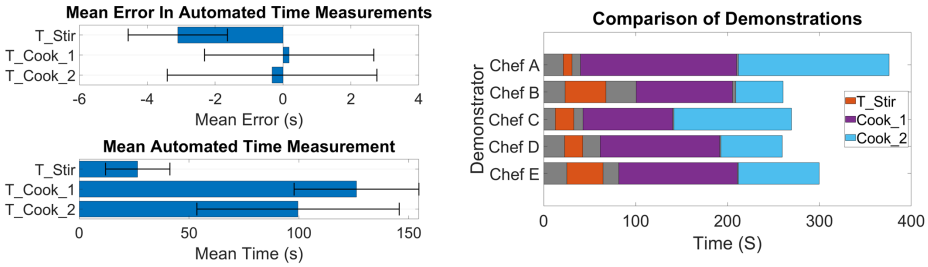


**Fig. 5.** Top left: mean error in automated measurements in seconds (+/- 1 standard deviation). The accuracy of the proposed timing extraction method was deemed acceptable for our purposes. Bottom left: mean automated time measurements across all demonstrators in seconds (+/- 1 standard deviation). Human cooking times significantly varied among different demonstrators. Right: compositions of full human cooking recipes. Substantial time is spent with the hands idle as the pancake cooked.

**Variation in Human Technique** The plots in Fig. 5 right and the large standard deviation bars in Fig. 5 bottom left show clearly that there were significant variations between the techniques employed by each demonstrator. Note in Fig 4 right that a significant amount of the time in a recipe demonstration involved the hands being idle as the pancake cooked. This idle time is part of the T_Cook_1 and T_Cook_2 times. Such idle time is interesting from the point of view of robotic automation of chores, as it has the potential to relieve a human chef of a time-consuming aspect of cooking, freeing up time that could be spent on other tasks. The relative range of values was largest for T_Stir, for which there was a factor of 4 between the shortest (11 s) and longest (49 s) recorded time. This demonstrates that there is utility in evaluating how the variations in these parameters affect the final quality scores.

**Pancake Quality Scoring** Visually, the pancakes produced by the robot seemed similar to those produced by the demonstrators, as shown in Figure 6 top right. This was matched by the quantitative results, with the pancakes produced by the UR5 found to be only marginally inferior to those produced by the human demonstrators. All quality categories had an average value between 4 and 5 (where 5 was the value assigned to the demonstrator's pancake). This suggests that our tracked key parameters encompass the most important aspects of the recipe. The average scoring of each of the 5 tested pancake pairs in each category is shown in Figure 6 bottom left.

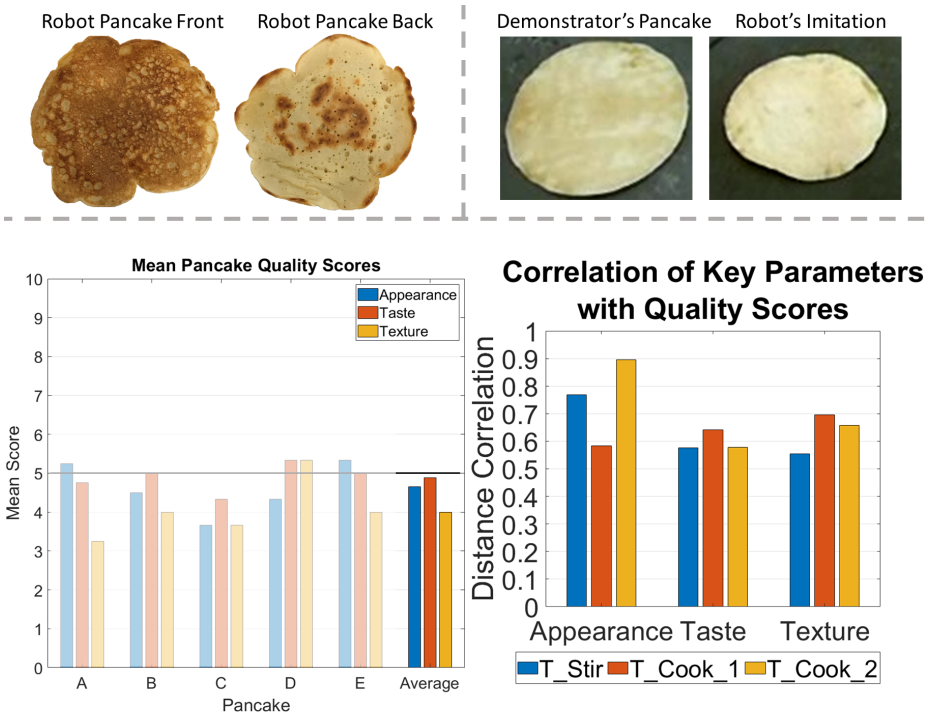


**Fig. 6.** Top left: images of the top and bottom of an example pancake cooked by the robot. Top right: a pancake cooked by a demonstrator compared to the corresponding pancake cooked by the UR5. Bottom left: robot pancake quality scores (human scores set to 5-5-5). A-E denotes the demonstrator imitated. Bottom right: distance correlation between key parameters and quality scores.

**Effect of Important Parameters** Distance correlation coefficients [12] were used to determine if any correlations existed between the quality scores and key times. The coefficients range from 0 to 1, 0 implying independence and 1

implying strictly obeyed linear correlation. In our scoring, we measured relative pancake quality scores, i.e. each robot-made pancake was scored relative to its human-made counterpart, not on a common scale, and aimed to exactly imitate the key parameters of the recipe. Hence ideally there would be no correlation between the difference in human and robot pancake quality scores and the key timings measured, as this would suggest that we had imitated the key times exactly. However, as can be seen in Fig 6 bottom right, some high magnitude correlations were observed.

Inaccuracies in our timing methods could have contributed to the correlations seen. The relative magnitude of the timing errors generally shrunk as the length of the time measured grew, which would likely have resulted in robot pancake quality scores closer to those of the human-made pancakes (i.e. 5). However, as seen in Figure 5 top left, the errors in the timing method were generally small and so are unlikely to be the sole cause of the correlations seen. Another possibility is that one of the recipe parameters deemed as not important (and so not tracked) is in fact important and is correlated with the key times, leading its effects to be represented in the figure. However, this is unlikely, as the only recipe parameter not tracked that could be potentially important is the manner in which the batter is poured on the hotplate, and this is not expected to be correlated with any of the key times tracked.

It is likely that a large portion of the correlations seen occur by chance in our small data set of 5 pancake pairs, and would reduce in magnitude as more datapoints are sampled. If this proves not to be the case, and larger datatsets reinforce the correlations shown here, this would suggest that the parameterisation method used did not completely encapsulate the effect of the measured parameters. Hence this style of analysis allows whether the recorded parameters are well tracked and replicated by the system to be determined. The current situation does not allow for gathering more data points, however we intend to resume the direction of research in the future.

## 5 Conclusion and Future Work

To address the problem of how unskilled end users could teach their recipe variants to domestic cooking robots, this paper has presented an initial investigation into a new methodology to enable such teaching, based on replicating important recipe parameters measured from human recipe demonstrations. Human demonstrators were recorded with a Kinect 2 sensor while carrying out a pancake recipe in a close to uninhibited manner. We developed a novel method to extract full 3D skeleton tracking from the single Kinect 2 sensor by combining the colour and depth videos. For this, the OpenPose neural network was used to extract 2D skeleton joint trajectories from the colour frames, which were then fused with the depth information of each joint to obtain the 3D trajectory reconstruction. The full pancake cooking process was split into timed stages, and we developed an algorithm to automatically extract the timings the human chefs spent on each stage using the 3D tracking trajectory. Three recipe stages were determined to

be crucial in replicating the pancake with a robotic chef. These were then passed to a UR5 to replicate as it carried out the recipe, using a mix of hardcoded pick-and-place sequences and tunable actions defined by the extracted timings. and the quality of the pairs of resulting pancakes was then evaluated.

Our goal in this work was to replicate a human-made pancake with a robotic chef such that they become indistinguishable for a human evaluator. Thus, the quality of the pairs of resulting pancakes was assessed through blind tests. The quality of the robot's pancakes was found to be just marginally inferior, suggesting that our chosen key parameters did encompass most of the important properties of the cooking process. A wide variety in key parameter values were observed, suggesting that there is utility in tracking and replicating these parameters. The relative influence of each tracked time on robot pancake quality (relative to that of the human pancake) was then evaluated, to determine if the system was well replicating the tracked parameters. High magnitude correlations were found, which would suggest that the quality difference between human and robot pancakes was dependent on the tracked timings that we attempted to replicate exactly. Such a result is surprising, and it is believed that this is most likely occurs by chance in our small sample size. Hence this approach for intuitively teaching recipe variants to robots is still viewed as promising, and we intend to resume the research in the near future.

This investigation has highlighted a large body of potential future work. Most importantly, further collection of data would clearly be advisable as it would allow more conclusive results to be drawn than what can be seen currently. This could also include a systematic variation of the key parameters to determine the sensitivity of the quality scores, e.g. if human testers could detect a difference in the quality scores if either of the time-sensitive stages is varied. Additionally, our current approach is scalable as long as the recipe processes are defined and robotic recipe execution can be divided into time-sensitive stages. We thus expect that the proposed method could readily be used to also reproduce other (relatively) simple recipes, such as the cooking of e.g. scrambled eggs, omelettes, or beans.

The current system assumes that a hand passing by a tool holder's location goes on to pick up said tool. While this has not lead to any errors in determining the recipe stage timings thus far, this method could be made more robust by incorporating computer vision to track tool positions in the colour images and determine if the tools were moving with the demonstrators hand. This could be extended by incorporating the Kinect's depth data to also determine the tool's pose. This would allow the recording of more sophisticated recipe parameters than the time spent on each recipe stage, and so open the door to better replication of the dishes produced by human chefs.

## 6 Acknowledgement

This work was supported by BEKO PLC and Symphony Kitchens, and the Engineering and Physical Sciences Research Council (EPSRC) RoboPatient grant [EP/T00603X/1].

## References

[1] Mario Bollini, Jennifer Barry, and Daniela Rus. "Bakebot: Baking cookies with the pr2". In: *The PR2 workshop: results, challenges and lessons learned in advancing robots with a common platform, IROS*. 2011.

[2] Michael Beetz, Ulrich Klank, Ingo Kresse, Alexis Maldonado, Lorenz Mösenlechner, Dejan Pangercic, Thomas Rühr, and Moritz Tenorth. "Robotic roommates making pancakes". In: *2011 11th IEEE-RAS International Conference on Humanoid Robots*. ISSN: 2164-0572. Oct. 2011, pp. 529–536.

[3] Moley Robotics. *Moley – The world's first robotic kitchen*. 2020. URL: https://www.moley.com (visited on 05/12/2020).

[4] Aude Billard, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal. "Handbook of robotics chapter 59: Robot programming by demonstration". In: *Handbook of Robotics. Springer* (2008).

[5] Riccardo Caccavale, Matteo Saveriano, Alberto Finzi, and Dongheui Lee. "Kinesthetic teaching and attentional supervision of structured tasks in human–robot interaction". en. In: *Autonomous Robots* 43.6 (Aug. 2019), pp. 1291–1307.

[6] Nadia Figueroa and Aude Billard. "Learning Complex Manipulation Tasks from Heterogeneous and Unstructured Demonstrations". en. In: *IROS Workshop on Synergies between Learning and Interaction* (2017), p. 7.

[7] Y. Mollard, T. Munzer, A. Baisero, M. Toussaint, and M. Lopes. "Robot programming from demonstration, feedback and transfer". In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Sept. 2015, pp. 1825–1831.

[8] Scott Niekum, Sarah Osentoski, George Konidaris, and Andrew G. Barto. "Learning and generalization of complex tasks from unstructured demonstrations". en. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Vilamoura-Algarve, Portugal: IEEE, Oct. 2012, pp. 5239–5246.

[9] Karinne Ramirez-Amaro, Michael Beetz, and Gordon Cheng. "Transferring skills to humanoid robots by extracting semantic representations from observations of human activities". In: *Artificial Intelligence* 247 (2017), pp. 95–118.

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: (2018).

[11] Abraham. Savitzky and M. J. E. Golay. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures". In: *Analytical Chemistry* 36.8 (1964), pp. 1627–1639.

[12] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. "Measuring and testing dependence by correlation of distances". In: *Ann. Statist.* 35.6 (Dec. 2007), pp. 2769–2794.