Origami Sensei: Mixed Reality AI-Assistant for Creative Tasks Using Hands

Qiyu Chen*
qiyuc@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, USA

Richa Mishra*
richamis@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, USA

Dina El-Zanfaly
delzanfa@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, USA

Kris Kitani
kmkitani@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, USA

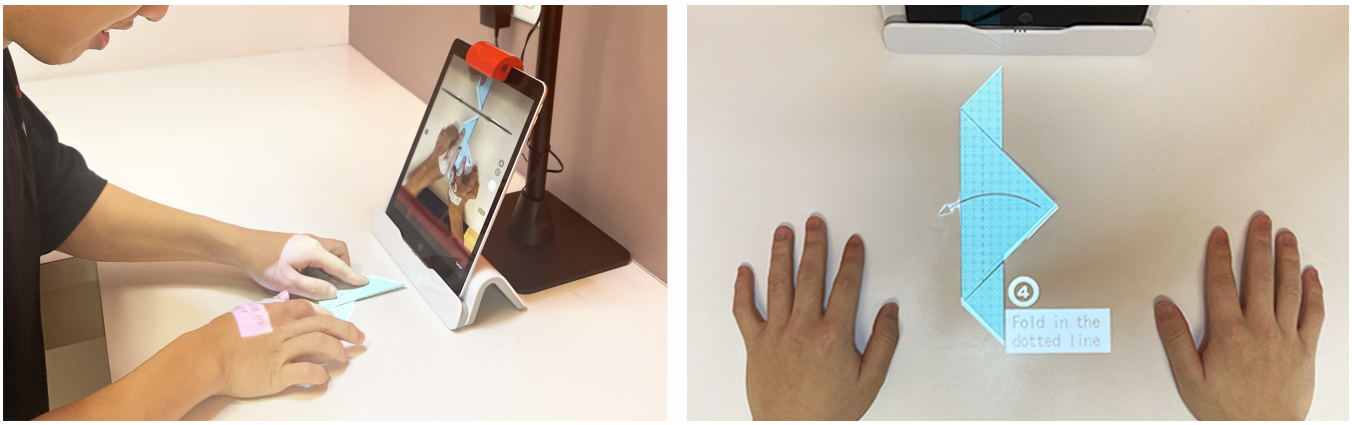

Figure 1: Origami Sensei, a mixed-reality system that assists beginners in creating origami by providing step-by-step instructions in real-time.

ABSTRACT

Learning creative tasks that involve using hands and the body, like knitting and music playing, can be challenging for beginners. We introduce Origami Sensei, a mixed-reality system that assists beginners in creating origami by providing step-by-step instructions in real-time. Origami Sensei enables users to create their own origami pieces from a library of pre-existing designs. It uses computer vision algorithms to identify the user's current origami step. It then provides them with real-time visual projections and verbal feedback to assist them in completing the next steps. The current setup consists of a tablet, a reflective mirror and a small projector. This research approach opens possibilities for introducing mixed-reality systems for learning hands-on skills in other fields such as welding, sculpting and even surgical operations. Future steps for this project include developing multimodal instructions and conducting user studies.

CCS CONCEPTS

• **Computer systems organization** → *Real-time system architecture*; • **Human-centered computing** → *Systems and tools for interaction design*; • **Computing methodologies** → **Activity recognition and understanding**; • **Applied computing** → **Fine arts**.

KEYWORDS

Convolutional neural networks, Data augmentation, PyTorch mobile, SwiftUI, Mixed reality, Computational making

ACM Reference Format:
Qiyu Chen, Richa Mishra, Dina El-Zanfaly, and Kris Kitani. 2023. Origami Sensei: Mixed Reality AI-Assistant for Creative Tasks Using Hands. In *Designing Interactive Systems Conference (DIS Companion '23), July 10–14, 2023, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3563703.3596625

1 INTRODUCTION

Learning creative tasks that involve using hands and the body, like origami, welding, or knitting, can be challenging for beginners. Traditional methods of instruction, such as instruction manuals

*Both authors contributed equally to this work.

DIS Companion '23, July 10–14, 2023, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9898-5/23/07.
https://doi.org/10.1145/3563703.3596625

or videos, do not provide feedback on learners' performance [20]. This feedback is essential in the learning process as it helps learners track their progress and identify areas for improvement [21]. By contrast, modern computer vision systems and mixed reality technologies can offer real-time feedback to users, allowing them to track their progress and correct errors as they occur. This approach is particularly helpful for beginners who need support in developing their skills. In this work, we focus on origami as a case study here for several reasons: First, compared to knitting and welding, origami is a simpler task with a standardized set of folding actions [17][13]; Second, origami requires only papers and a few basic tools, rendering it a low-cost and accessible task for people from all backgrounds [22]. Lastly, origami is a highly planar visual task, which makes it easier for a computer vision system to capture and provide real-time feedback.

Learning origami through traditional methods, like instructional manuals and videos, can be difficult for beginners [20][21]. These methods may lack detailed explanations, assume prior knowledge, use unclear jargon, and only present discrete steps and final results. Also, these instructional methods are limited to two dimensions, either printed paper or digital screens, and do not provide guidance on what to do in the event of mistakes.

We introduce Origami Sensei, a mixed-reality system that assists beginners in creating origami by providing step-by-step instructions in real-time. Origami Sensei enables users to create their own origami pieces from a library of pre-existing designs. It uses computer vision algorithms to identify the user's current origami step. It then provides them with real-time visual projections and verbal feedback to assist them in completing the next steps. The current setup consists of a tablet, a reflective mirror and a small projector. The complete system uses computer vision models and algorithms to achieve three primary functions: (1) recognize the current step of the user; (2) offer guidance on how to proceed to the next step; and (3) detect when the user makes a mistake and provide tailored instructions to rectify it. Compared to existing instructional methods, Origami Sensei offers a new modality of origami instruction with an unprecedented level of learner-centric interaction and automation, which means placing the learner's needs and interests at the center of the learning experience. This research approach opens possibilities for introducing mixed-reality systems for learning hands-on skills in other fields such as welding, sculpting and even surgical operations.

This work-in-progress paper presents our progress on developing the system. We begin by reviewing the background and related work on AI-assistant origami teaching and defining our new system setup. We then collect origami folding data, and achieve the first main functionality of step recognition by training convolutional neural network classifiers. Moving forward, we will implement additional computer vision algorithms to accomplish the remaining two functions of Origami Sensei and conduct a comprehensive survey to evaluate the system.

2 BACKGROUND AND RELATED WORK

2.1 Origami and Traditional Forms of Instruction

The word "origami" comes from the Japanese words "ori" (to fold) and "kami" (paper) [1]. The idea behind origami is simple: take a single sheet of paper and fold it sequentially into a beautiful, often intricate, artwork.

However, mastering the art of origami, or even simply following the instruction, is challenging due to several factors. For instance, some folds in the instruction can have multiple candidates from a fixed viewpoint [21], and certain inherently-3D folds like the closed sink fold are hard to represent in two dimensions [24]. Moreover, traditional forms of origami instruction lack the intuitiveness, convenience, and interactive elements offered by modern technology [22][20]. There are assorted traditional forms available, ranging from the least learner-centric, such as crease patterns (creases on an unfolded paper) [15], to the most learner-centric, like instructional videos, which record the entire folding process. Nevertheless, even the most detailed video cannot confirm if the last step is folded correctly or project instructions directly onto the paper.

Mixed reality AI-assistant forms of instructions have been extensively studied for other tasks like driving [6], motherboard assembly [18], and welding [2][11]. Given the growing importance of origami in fields like architecture, robotics, and astronautics [15], there is a need for a learner-centric mixed reality AI-assistant system for origami learning to make it easier and more accessible for learners to master complex folds involved in origami creation.

2.2 Existing Pipelines and Applications

Various approaches have been employed in developing a visual and digital origami teaching system. One simple approach is the use of static images depicting the output of each step along with textual instructions [24]. Users using this system would have to manually advance to the next state, and there is no feedback on users' progress. This kind of system is essentially a digital version of instruction manuals. Unfortunately, most origami teaching systems today are some version of this.

The most relevant prior pipeline to our work is from a series of studies conducted by a group of Japanese researchers from 1997 to 2020 [19][12][23][20][21]. These studies also aimed to automatically recognize origami steps and provide feedback to users. However, their pipeline relied heavily on machine learning algorithms and involved numerous pre-processing and post-processing steps, making it overly complicated for developers. In addition, their approach was limited to the set-up of a white paper on a black background, and they restricted the definition of each step to only the silhouette of the paper, without taking into account details like creases inside the paper. Lastly, their approach to instruction was limited to a computer application that displayed the step recognition results and instructions for the next steps on the screen only.

Recent work [22] developed a mixed reality system to aid origami learning. However, their work suffered from minimal interaction and a lack of automated step tracking and tailored instructions. For instance, users must manually click a button to see the instruction for the next step. While previous works, such as [16] and the

PlayGAMI application [8], have achieved real-time automatic paper or step tracking, they depend on visual cues on specialized paper surfaces, which are not used in common origami papers and thus not widely applicable.

By contrast, the neural networks in our method can leverage the inherent patterns underneath the data to achieve automatic step recognition based solely on the appearance of the paper without making any assumptions about the appearance of the background or the paper itself. The complete pipeline can provide instructions through projection and guide users to correct mistakes as well.

3 ORIGAMI SENSEI

The current setup consists of a tablet, a reflective mirror and a small projector. We use the tablet's camera to automatically track the progress of the user in real time and utilize a projector to provide folding instructions overlaid onto the tracked paper. In order to achieve this, we exploit data-driven models to replace some machine learning algorithms used in the previous pipeline [21] which engenders restrictive set-up, overly simplified assumptions, and complicated data processing steps.

In our design, depicted on Fig. 2, an iPad is used along with a mirror attachment from Osmo [10] attached to its front-facing camera. This configuration enables the system to capture video of the user building origami while simultaneously displaying visual guidance on the front screen. From RGB image input, the system predicts the current state of the origami being constructed. When the user successfully arrives at a new step, the system will provide instructions for the next step on the screen and projection on the physical origami paper. However, if a mistake is detected, the system offers tailored feedback to help the user correct the error. During actual usage, when users finish a step, they move their hands away from the paper (since it will be predicted as the transition state otherwise), and immediately receive feedback based on the model prediction. The system gives feedback by overlaying folding instructions on top of the physical paper using a small projector, along with some optional text. The whole process is repeated until the user attains the last step. This pipeline is also shown in Fig. 3.

3.1 Step Recognition Classifier

The main functionality of our vision based system is to automatically recognize the user's current step in real-time. We view this step recognition problem as a multi-class classification task tackled using convolutional neural networks. Each state, either a completed step or a transition state in between two steps, of the origami paper will be predicted by the convolutional neural network classifier. For the sake of accuracy, we define every image frame with hands occluding the origami paper as the transition state. Such design choice largely removes the adverse effect of hand occlusion during step recognition without enforcing large demands on the user's side.

We initialize a ResNet18 model [9] with weights pre-trained on ImageNet [4] and replace the last classification layer with our own. Then we fine-tune the whole model using our dataset explained in Section 3.2. Currently, we train one model per origami design, since we want to focus on optimizing the instruction-giving module rather than generalizing over many origami models.

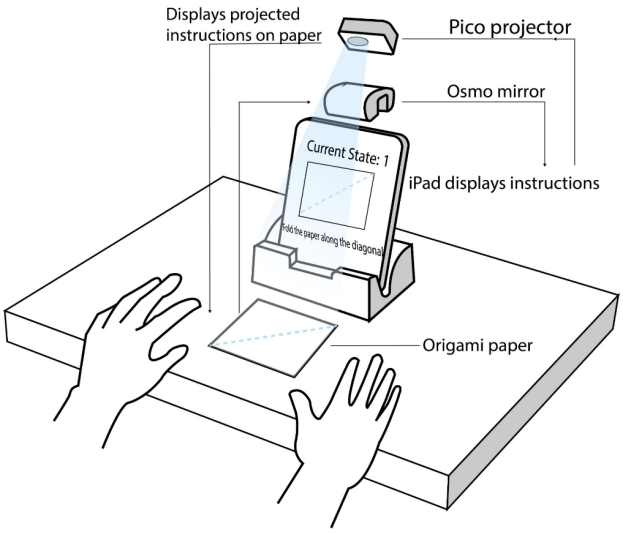

Figure 2: Our system design diagram.

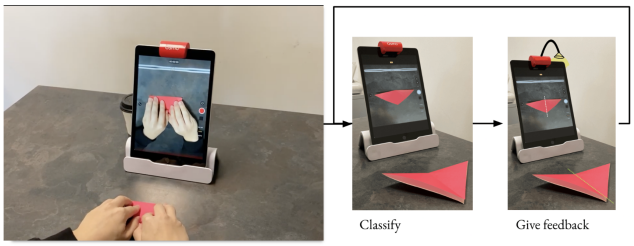

Figure 3: Current origami state estimation pipeline.

3.2 Dataset

To develop robust computer vision models that can effectively recognize different origami states, we need datasets for model training. An alternative perspective to view this is that, [21] manually documents all the operation details of every step of the folding process into specific data structures, while we use data-driven neural networks to implicitly learn the visual characteristics of each fold. Another benefit of using neural networks trained on datasets is that we can achieve generalization across various paper types and background patterns through data augmentation.

Although there are a few publicly available origami datasets [14], they are not appropriate for step recognition. Therefore, we gathered our own dataset by recording videos of an individual creating origami using both an iPad with an Osmo mirror attached (as shown in Fig. 1) and a cellphone positioned above the workspace. It is worth noting that the captured images are consistent in terms of distance to the paper and lighting to capture maximum amount of details. Additionally, we pre-process the videos by extracting frames at 10 frames per second, cropping the area of interest, and applying a homography transformation to correct for the perspective distortion caused by the Osmo mirror. In the end, we collect over 23,000 image frames. Since all transition states are pooled together to state 0, they easily dominate the dataset. To address

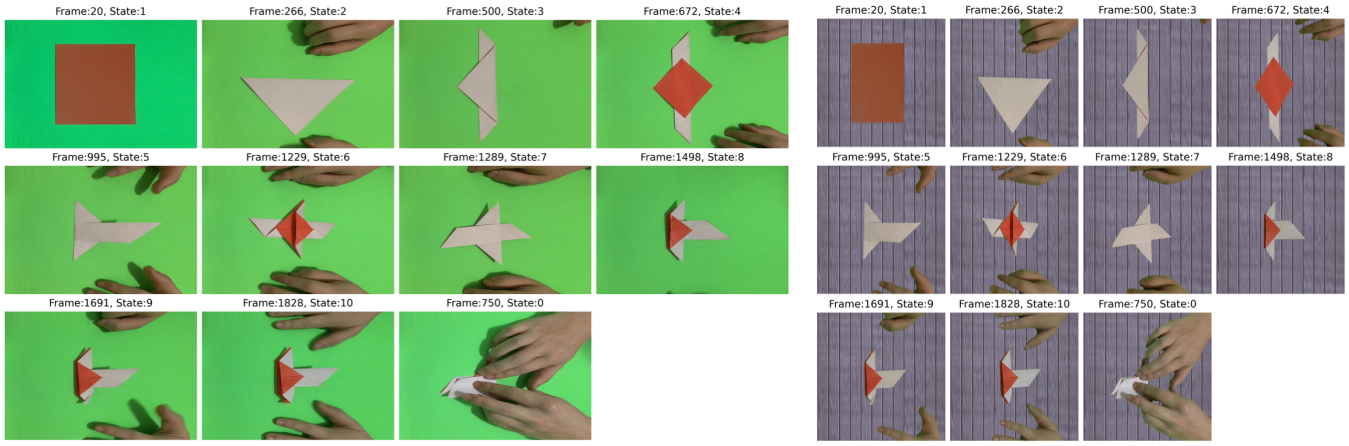

Figure 4: Our dataset for an origami dove. Left: examples of pre-processed images for each state. Right: the same set of images after resizing and background replacement.

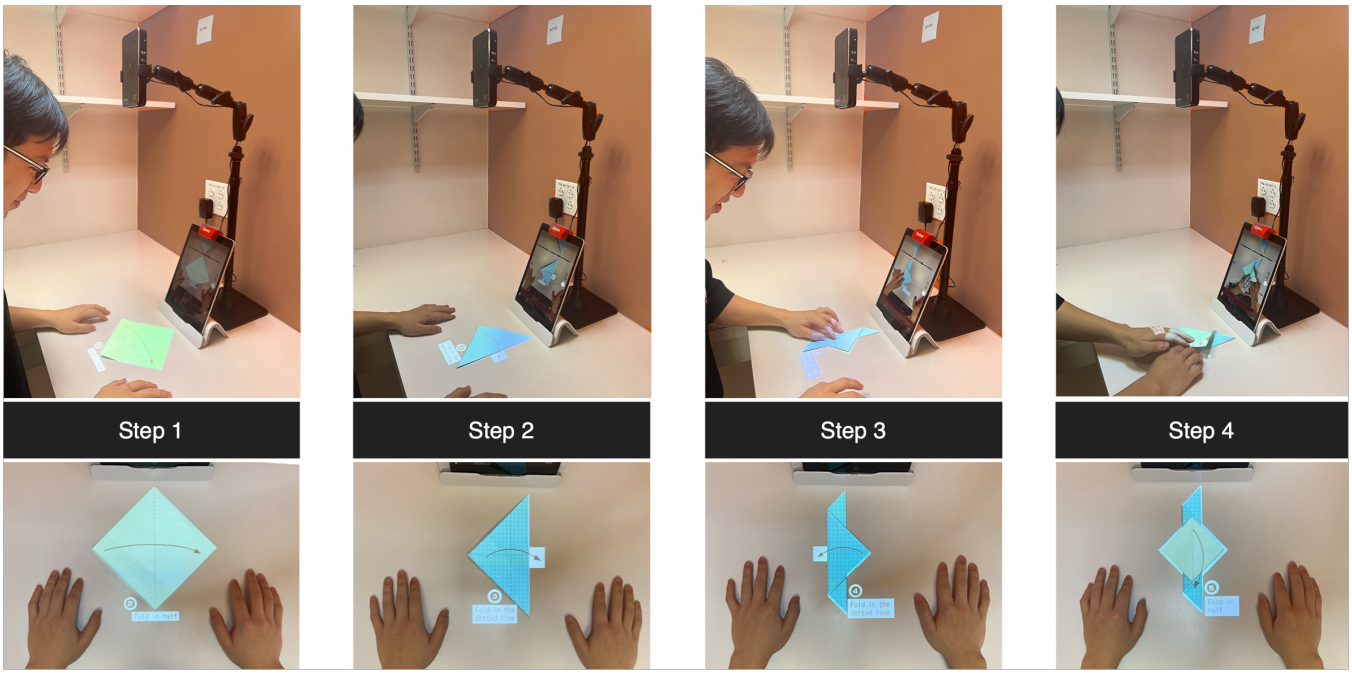

Figure 5: Demonstration of Origami Sensei overlaying instructions onto the physical origami paper for steps 1, 2, 3, and 4 of an origami dove model. The top row shows the whole system and a user following the instructions to fold the paper. The bottom row provides close-ups of the user's view of the projected instructions and descriptions.

this class imbalance issue, we randomly select samples from each state based on the minimum frequency of all states. The resultant samples for an origami dove model are displayed in Fig. 4 left.

To increase data efficiency and avoid collecting and manually annotating hundreds of videos, we augment our data via green screen techniques on the five collected videos to replace the background with other texture images from [5][7][3]. Specifically, we synthesize 10 new videos with different texture background (Fig. 4 right). Furthermore, during training, we also employ online data augmentation techniques including flipping, rotation, color jittering, perspective distortion, etc. so that the trained network can generalize to a reasonable spectrum of background texture, paper colors, lighting, distance, and angles. The final dataset contains 7,920 images (720 images per state) for training and 880 images for testing. Since these test samples are from the same videos as training samples, they differ only in the background pattern and the

effect caused by online data augmentation. In order to estimate the model's out-of-distribution (OOD) generalization, we collect a new set of 660 test images (60 images per state) from videos recorded under different lighting using unseen paper colors and table patterns across 360 degree rotation and different distance.

4 EXPERIMENTS

Our classification model achieves a high accuracy of 97.8% on the normal test set. When tested on OOD data recorded under different set-ups and angles, the accuracy drops to around 75%. We believe such performance drop can be mitigated by incorporating more data or heavier data augmentation. It is also possible to finetune the model by taking a few images in the new background before running inference. To the best of our knowledge, our pipeline is the first to use the convolutional neural network on origami step classification, and our experiment proves its ability to implicitly learn the characteristics of different steps. Currently, we are developing an iOS app using PyTorch Mobile and SwiftUI to deploy the network on an iPad in order to test the generalization ability in real time on different origami models.

5 CONCLUSION AND FUTURE WORK

We aim to improve hands-on skills for creative tasks through the use of mixed-reality and computer vision systems. We are currently developing Origami Sensei, a mixed-reality system that helps beginners create origami pieces by providing real-time step-by-step projected instructions. Our approach involves using computer vision algorithms to identify the user's current origami step and to provide real-time visual projections and verbal feedback to assist learners. We show that data-driven models that we developed can predict the current state of the origami accurately. After deploying the trained model on iPad, we will iteratively add and test more advanced features towards the complete system pipeline.

5.1 Designing the Instructions

For next steps, we will co-design the verbal and projected instructions with users. We are interested in how the instructions will be overlaid on the physical origami paper. Additionally, we aim to determine what level of visual and verbal guidance that origami Sensei should provide. We will also add a "mistake" state in the classification model and incorporate a Hidden Markov Model on top of it to add another level of control and predict users' mistakes. Based on the prediction, using the projection interface and another paper pose estimation model, the system can give continuous feedback by overlaying instructions onto the physical paper that the user is folding, while simultaneously showing the instructions on the iPad screen, as illustrated in Fig. 5.

5.2 User Study

We will develop and evaluate our system with users of different skill levels in origami, as well as with origami models of varying levels of difficulty. We aim to gather user feedback of the system's effectiveness in providing a learner-centric and interactive origami learning experience.

REFERENCES

[1] 2021. The origin of origami: What events shaped this ancient papercraft? *How It Works Magazine* 152 (June 2021), 66.

[2] Richa Agrawal and Jayesh S. Pillai. 2020. Augmented Reality Application in Vocational Education: A Case of Welding Training. *Companion Proceedings of the 2020 Conference on Interactive Surfaces and Spaces* (2020).

[3] Gertjan Burghouts and Jan-Mark Geusebroek. 2009. Material-specific adaptation of color invariant features. *Pattern Recognition Letters* 30 (February 2009), 306–313. https://doi.org/10.1016/j.patrec.2008.10.005

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[5] Michal Franczak. 1993. Download free classic textures from Pixar. https://evermotion.org/articles/show/9410/download-free-classic-textures-from-pixar

[6] David Goedicke, Alexandra W. D. Bremers, Sam Lee, Fanjun Bu, Hiroshi Yasuda, and Wendy Ju. 2022. XR-OOM: MiXed Reality driving simulation with real cars for research and design. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[7] Chris Graczyk. 1995. VisTex database. https://vismod.media.mit.edu/vismod/imagery/VisionTexture/vistex.html

[8] Uttam Grandhi and Ina Yosun Chang. 2019. PlayGAMI: augmented reality origami creativity platform. *ACM SIGGRAPH 2019 Appy Hour* (2019).

[9] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 770–778.

[10] Simon Hill. 2015. Osmo review. https://www.digitaltrends.com/gaming/osmo-review/

[11] Ananya Ipsita, Levi Erickson, Yangzi Dong, Joey Huang, Alexa Bushinski, Sraven Saradhi, Ana M. Villanueva, Kylie A. Peppler, Thomas S. Redick, and Karthik Ramani. 2022. Towards Modeling of Virtual Reality Welding Simulators to Promote Accessible and Scalable Training. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[12] Jien Kato, Hiroshi Shimanuki, and Toyohide Watanabe. 2006. Understanding and Reconstruction of Folding Process Explained by Illustrations of Origami Drill Books.

[13] Robert J. Lang. 2011. ORIGAMI DIAGRAMMING CONVENTIONS. https://langorigami.com/article/origami-diagramming-conventions

[14] Daniel Ma, Gerald Friedland, and Mario Michael Krell. 2021. OrigamiSet1.0: Two New Datasets for Origami Classification and Difficulty Estimation. *CoRR* abs/2101.05470 (2021). https://arxiv.org/abs/2101.05470

[15] Marco Meloni, Jianguo Cai, Qian Zhang, Daniel Sang-Hoon Lee, Meng Li, Ma Ruijun, Teo Parashkevov, and Jian Feng. 2021. Engineering Origami: A Comprehensive Review of Recent Applications, Design Methods, and Tools. *Advanced Science* 8 (May 2021). https://doi.org/10.1002/advs.202000636

[16] Atsushi Nakano, Makoto Oka, and Hirohiko Mori. 2013. Leaning Origami Using 3D Mixed Reality Technique. In *Human Interface and the Management of Information. Information and Interaction Design*, Sakae Yamamoto (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 126–132.

[17] Samuel Randlett. 1961. The Art of Origami; Paper Folding, Traditional and Modern.

[18] Arvin Christopher C. Reyes, Neil Patrick Del Gallego, and Jordan Aiko Deja. 2020. Mixed Reality Guidance System for Motherboard Assembly Using Tangible Augmented Reality. *Proceedings of the 2020 4th International Conference on Virtual and Augmented Reality Simulations* (2020).

[19] Hiroshi Shimanuki, Jien Kato, and Toyohide Watanabe. 2002. Constituting Feasible Folding Operation Using Incomplete Crease Information. In *IAPR International Workshop on Machine Vision Applications*.

[20] Hiroshi Shimanuki, Yasuhiro Kinoshita, Toyohide Watanabe, Koichi Asakura, and Hideki Sato. 2016. Operational Support for Origami Beginners by Correcting Mistakes. *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication* (2016).

[21] Hiroshi Shimanuki, Toyohide Watanabe, Koichi Asakura, Hideki Sato, and Taketoshi Ushiama. 2020. Anomaly Detection of Folding Operations for Origami Instruction with Single Camera. *IEICE Trans. Inf. Syst.* 103-D (2020), 1088–1098.

[22] Yingjie Song, Chenglei Yang, Wei Gai, Yulong Bian, and Juan Liu. 2020. A new storytelling genre: combining handicraft elements and storytelling via mixed reality technology. *The Visual Computer* 36 (2020), 2079 – 2090.

[23] Toyohide Watanabe and Yasuhiro Kinoshita. 2012. Folding support for beginners based on state estimation of origami. In *TENCON 2012 IEEE Region 10 Conference*. IEEE, 1–6.

[24] Guy Zimmerman, Julie Barnes, and Laura Leventhal. 2003. A Comparison of the Usability and Effectiveness of Web-Based Delivery of Instructions for Inherently-3D Construction Tasks on Handheld and Desktop Computers. In *Proceedings of the Eighth International Conference on 3D Web Technology* (Saint Malo, France) *(Web3D '03)*. Association for Computing Machinery, New York, NY, USA, 49–54. https://doi.org/10.1145/636593.636601