(2022), **1**, 1–4

Fingertip detection and locus tracking for air-drawn sketches using only video data

Dibyanshu Shekhar,* Millennium Bismay,* and Sambhav Khurana*

*Email: dshekhar@tamu.edu; mbismay@tamu.edu; sambhav_khurana@tamu.edu

Abstract

Air sketching is the process of drawing/sketching in air using finger and movements or gestures without the use of any sketching aid or reproducible physical surface. Even after significant advances in object detection and object tracking, analogous analysis of air sketches using finger tracking has been a challenging task. There has been some advancement in the recognition of single stroke segments but, analysis on multi-stroke shapes/sketches is wanting. We propose a novel method of fingertip detection and locus tracking using YOLO-v4 architecture as the backbone for object detection in Realtime, segmentation of collated hand contours, evaluation of finger point region and tracking and the constant scene delimiter, a unique proposition to tackle multi-stroke sketches. This method is expected to provide an extremely efficient detection of air-drawn sketches, not just limiting to single stroke detection but effectually extending to multi-stroke sketch analysis.

Keywords: Object detection, finger tip tracking, constant scene delimiter

1. Introduction

Modern day sketch recognition and vision tasks mostly are operated on images of drawn sketches, which essentially are reproduced or collected using physical devices such as kinect or on physical drawing boards. As the world is gearing towards Human Computer Interaction, with Facebook showcasing Oculus VR headsets, google with its AR glasses, Mircrosoft HOLOLens etc. sketching should also propagate to such a dimension. There has been significant work in vision based systems with regard to air-writing based recognition, but most of them use external drawing aids or detection sensors to collect data from users. In the Air-writing paper by Amma C., inertial sensors gloves were used for continuous spotting and recognition of air-writing. Another work, more into the advances in sensor technology, by Chang includes the usage of depth sensors such as Microsoft Kinect and specialized hardware such as leap motion controller for input.

Usage of these hardware peripherals tend to distort user information collected during experiments, as user is more conscious of his/her surrounding, less indicative of real world scenarios of free drawing and hand gesture movements. Thus it is essential to free the extra burden of wearing and handling extra equipment to emulate real-life air sketching i.e. to move towards a more convenient and less cumbersome approach, analyzing video data.

Huang had proposed a two stage CNN-based fingertip detection framework for recognition of air-writing in egocentric(first person) RGB video. However, capturing egocentric video requires head-mounted smart cameras or mixed reality headsets which may not be available to all users. Thus, to make our proposed system more ubiquitous, we use video inputs from a standard laptop camera, web-cam or mobile phone camera for the air-sketching application. This makes the task even more

daunting due to the presence of the face and other prime landmarks in video frames which might delude proper hand and subsequent finger detection.

2. Related Work

3D hand-tracking was proposed by Tang, which required very high computational cost as well as large amount of training data, thus is not a good fit for realtime finger tracking. Non-intelligence based methodologies include algorithms to segment hand contours using color, depth and motion cues and extract binary masks. Liang proposed a distance metric from hand palm to the contour furthest points to localize candidate fingertip points and appropriate prediction from these selections. These methods suffer significantly when hand segmentation was executed poorly. Recent advancements in segmentation using deep learning based approaches, particulary the U-net architecture has contributed significantly to predict precise segmentation silhouttes. PSPNet is another such fast and precise architecture for image segmentation even on multi-class domain.

Faster R-CNN(Region-based Convolutional Neural Networks) is one of the promising methods for object detection, as it takes region based contexts and forms a proposal network to identify objects in an image. Roy proposed a two-step framework for detection and segmentation of hands using a Faster R-CNN based hand detector followed by a CNN based skin detection technique. Wu et proposed a framework called YOLSE (You Only Look what You Should See), that uses a heatmap-based fully convolution network for multiple fingertip detection from single RGB images in egocentric view. This is an interesting concept which can be extended to time series domain , to evaluate on video data. Mayol et. al. (and Kurata have used template matching and mean-shift respectively for hand tracking in constrained environments, from images captured using a wearable camera. These methods can protract to long term video sequences as these are designed for short videos wih limited fps, as inefficient methods such as template matching will lead to extreme lag in detection. Closest to the work we are presenting is the paper by Sohom et al. in which they use faster RCNN based approach for object detection, a velocity metric to provide delimiters for start and end of writing/sketch and smoothening algorithm to fix small disturbances which may be caused due to handle trembling or any other noise inducing factor.

Our proposed solution tackles many of the above issues mentioned that include precise hand segmentation and detection with U-net and YOLOv4 architecures respectively, providing real-time performance with efficient computation facilitating a low memory footprint. One of the major factors overlooked by previous solutions is the reckoning of multiple strokes in an air sketch, which is very common in intricate shapes and designs.

3. Proposed Methodology

We propose novel methodologies for pointer detection and scene delimeter which are crucial to develop a multi-stroke air sketch techniques. The complete flowchart is presented in Figure 1.

3.1 Palm Region Detection

The detection of writing palm is an integral part of air sketching. We require fast and efficient detection of the palm region from the video data to localize the region, which is essential for locating the drawing device, which is the finger pointer. We propose using YOLOv4 and MobilenetV2 in our initial iterations. YOLOv4 is known to be more precise but slower as compared to MobilenetV2. Hence, depending on the frames required for efficient tracking and detection of intricate shapes, we will finalize the palm detection algorithm.

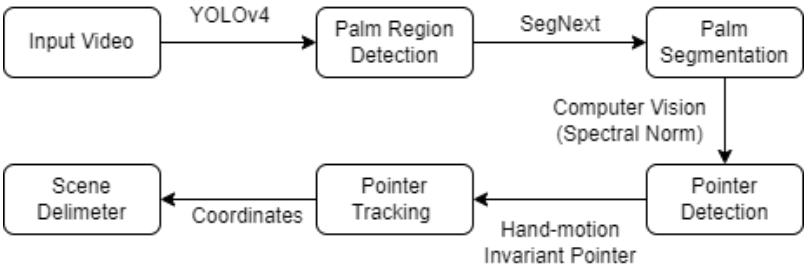

Figure 1. Proposed Methodology

3.2 Palm Segmentation

Once, the palm is precisely detected, we propose using SegNext for semantic segmentation of the palm from the background. SegNext uses convolutional attention to encode contextual information. Palm segmentation enables us to remove any noise which could distort the shape of the figures drawn during air-sketching by focusing only on the palm region and helps in its efficient tracking.

3.3 Pointer Detection

Unlike traditional pen drawing where the hand rests on a surface, the hand or the fingertip hovers in the air during air-sketch without any form of support. This creates problems for users while drawing multiple strokes or long strokes and during the transition from one shape to another. We propose a novel solution to tackle this problem by introducing the concept of *Pointer*. A *pointer* is defined as the farthest point from the centroid of the palm. From the precise segmentation from the above technique, we will derive the centroid of the palm and will implement the Spectral Norm on the segmented boundary to detect the pointer. This will make the air-sketching robust and invariant to hand rotations and multiple finger movements during air-sketching.

3.4 Pointer Tracking

We will be precisely tracking the pointer, defined in our previous step, to find the spatial coordinates of the air-sketch. The coordinates are simply present in the hovering space of the pointer. The coordinates need to be precise for the accurate detection of intricate shapes.

3.5 Constant Scene Delimiter

This segment of the system will act as the delimiter for the start and end of a sketch. This will be a part of the real-time detection system, analyzing movement(essentially hand position detection and evaluation) in consecutive frames and stopping the frame processing once the scene doesn't change for a heuristically determined threshold number of frames. Most multi-stroke sketches have the strokes drawn withinin 4-5 frame buffers within each stroke, no change in existing pointer position for more than the threshold number of frames will indicate the end of the sketch.

3.6 Research Questions

1. How to detect sketch initiation without any definite writing hand pose or orientation?
2. How to account for multiple strokes in a complex sketch?
3. What other factors can be taken into consideration to identify end of a sketch, apart from the velocity metrics?
4. How to devise an algorithm to make air-sketching invariant to hand rotation and multiple finger movements?

4. Evaluation: Analysis Plan

For hand detection and tracking YOLOv4 is to be trained on 15000 images available from EgoFinger dataset and the EgoHands dataset with a backbone of bottleneck Resnets or EfficentNet model architectures. The evaluated frame regions will be used to segment foreground hand information from background using SegNext or ViT(Vision Transformers) based architectures which are the state-of-the art in image segmentation. Videos will be collected from phone cameras, pre-pocessing and down-sampling on per-frame basis will be performed for efficient and faster training. Due to high FPS rates in phone cameras, interpolation methods will be use to drop sufficiently identical frames to improve computational efficiency and simultaneously remove redundancy. The constant scene delimiter will keep track of incoming frames and stop processing on receives greater number of frames than the heuristically evaluated threshold. The tracked coordinates calculated at each pointer detection will be mapped to screen coordinated to project the air-sketched shape.

5. Evaluation: Definition of Success

The proposed system should provide near-precise air-drawn sketch detections supporting multiple strokes by the user. It'll advance towards a better medium of pictorial communication, making the air sketch recognition, hassle free and independent of heavy-duty equipments and gadgets.

6. References

- Amma, C., Georgi, M., and Schultz, T. (2012). Airwriting: Hands-free mobile text input by spotting and continuous recognition of 3d-space handwriting with inertial sensors. In Wearable Computers (ISWC), 2012 16th International Symposium on, pages 52–59. IEEE.
- Behera, S. K., Dogra, D. P., and Roy, P. P. (2017). Analysis of 3d signatures recorded using leap motion sensor. Multimedia Tools and Applications
- Chang, H. J., Garcia-Hernando, G., Tang, D., and Kim, T.-K. (2016). Spatio-temporal hough forest for efficient detection–localisation–recognition of fingerwriting in egocentric camera. Computer Vision and Image Understanding
- Chen, M., AlRegib, G., and Juang, B.-H. (2016). Air-writing recognitionpart ii: Detection and recognition of writing activity in continuous stream of motion data. IEEE Transactions on Human-Machine Systems
- Deng, X., Zhang, Y., Yang, S., Tan, P., Chang, L., Yuan, Y., and Wang, H. (2018). Joint hand detection and rotation estimation using cnn. IEEE Transactions on Image Processing
- Liang, H., Yuan, J., and Thalmann, D. (2012). 3d fingertip and palm tracking in depth image sequences. In Proceedings of the 20th ACM International Conference on Multimedia