On De-Interlacing and Sub-Pixel Precision Tracking

Nawaz Aakif, Kuronen Toni, Eerola Tuomas, Lensu Lasse, Kälviäinen Heikki

DOI: 10.1109/IVCNZ54163.2021.9653265

On De-Interlacing and Sub-Pixel Precision Tracking

Aakif Nawaz*†, Toni Kuronen†, Tuomas Eerola†, Lasse Lensu† and Heikki Kälviäinen†
*Huawei Technologies Oy, Tampere, Finland
Email: aakif.nawaz@huawei.com
† Computer Vision and Pattern Recognition Laboratory, School of Engineering Science
Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland
Email: firstname.lastname@lut.fi

***Abstract*—Video cameras with interlaced scan sensors find applications in a variety of tasks such as object tracking due to their lower overhead in terms of memory and the higher sensitivity in comparison to their counterparts that employ progressive scan sensors. Such cameras, however, suffer from noticeable interlacing artefacts that need to be corrected with appropriate de-interlacing methods before the target in the video can be accurately tracked. Despite this, the effect of de-interlacing methods on the object tracking accuracy has not yet been widely studied. In this work, the first comprehensive comparison of different de-interlacing methods is carried out in the context human computer interaction studies where precise finger tracking is required. Furthermore, we propose a semi-automatic sub-pixel annotation scheme to create precise ground truth for fingertip location, allowing the analysis of the impact of de-interlacing filters on tracking at sub-pixel level. The experimental part of the work showed that the de-interlacing filter by Pixop outperformed other filters that were evaluated. Moreover, the plausible benefits of sub-pixel precise tracking over pixel precise tracking in trajectory analysis were demonstrated.**

***Index Terms*—De-interlacing, Hand Tracking, Key-point Annotation, Human Computer Interaction**

I. Introduction

Human-Computer Interaction (HCI) is an endeavour that seeks to understand and transform our interactions with complex technological artefacts in order to make these interactions effective, efficient and more importantly, enjoyable. With the recent advancements in gesture recognition, touch screens, augmented and virtual reality, studies that seek to better understand potential usability concerns have gained prominence. While data gloves that employ electro-mechanical, infrared or magnetic sensors offer a robust avenue to address the challenge of accurately tracking hand movements in HCI studies [1], these devices, however, hinder natural hand movements and are hence not suited for such usability studies. Commercial computer vision based technologies such as Leap Motion sensor [2] and Microsoft Kinect™ on the other hand can capture such interactions in an unimpeded manner, but are limited by their narrow range and low frame rate respectively.

High speed cameras, that can capture video at frame rates exceeding 250 frames per second (fps), coupled with visual object trackers have since emerged as a viable solution [3], [4]. Such a setup can be augmented inexpensively, with the aid of a normal speed camera to infer 3-D real world coordinates [5]. However, a minor blemish associated with such a framework emanates due to the inability of the generic object trackers to accurately localize the hand and finger movements to the required level of precision leading to a significant number of overshooting errors [6]. This apathy of object trackers towards sub-pixel accuracy is warranted as most object tracking algorithms are geared towards tasks that consider robustness and real time processing, over accuracy, as more critical factors in evaluating the success of a tracking algorithm. However, in touch screen usability studies, sub-pixel accuracy takes precedence as even minor errors in target localization can lead to huge errors, especially while considering parallax errors [7].

978-1-6654-0645-1/21/$31.00 ©2021 IEEE

An interesting aspect that is commonly ignored in similar usability studies is the fact that a large percentage of computer vision applications rely on cameras that employ interlaced scan sensors. The reason behind this is two fold: low overhead related to memory and the higher sensitivity of interlaced camera sensors compared with cameras having progressive scan sensors [8]. However, in the event that an object in a given frame exhibits considerable motion between consecutive frames, tearing or interlacing artefacts are induced near regions surrounding such objects which can significantly hinder the applicability of visual object trackers. De-interlacing algorithms that interpolate each half-frame into a corresponding full frame have emerged as a possible approach to handle such artefacts while also doubling the number of frames. This, however, is a fundamentally impossible task and brings about considerable degradation to the quality of the video frames.

In this paper, efforts have been directed towards refining the previously concluded usability study on 3-D touch screens by Kuronen *et al.* [9] in the context of the frailties discussed above. Specifically, we compare the various de-interlacing filters, including a recent deep learning based de-interlacer [10] and their implications on the performance of the visual object trackers. Moreover, the benefits of sub-pixel precise tracking over pixel precise tracking is studied by evaluating the velocity and acceleration curves of the tracked fingertips. To enable this, a robust sub-pixel annotation scheme is proposed to allow for the creation of high precision ground truth for tracker evaluation.

II. Related Work

A. Visual Object Tracking

Visual object tracking is amongst the most rapidly developing field in computer vision with applications in various disciplines such as robotics, automation and surveillance systems. The main objective of this field is to reliably estimate the state of a certain target object across all the frames of any given video sequence, given its state in the first frame [11]. The state of an object refers to information pertaining to the object's position, appearance and shape. An object tracking framework essentially comprises three core elements: an object representation model, a dynamic model, and a search mechanism [12], of which the object representation model has attracted an overriding amount of research interest and forms the basis of classification of object trackers into generative and discriminative methods [13]. Generative methods concentrate on finding areas within a frame that are congruent to the target object. On the other hand, discriminative methods distinguish the target objects from the surrounding background and essentially approach object tracking as a classification problem.

Previously, object trackers were known to be inept at their handling of various challenges posed by partial occlusions, motion blurs, scale variations and illumination changes, among others, resulting in considerable research effort being directed to this cause. These shortcomings, however, have been addressed considerably since the inception of the annual Visual Object Tracking (VOT) Challenge [14] that made it possible to achieve performance standardisation in tracker evaluation over its carefully compiled data-sets. Over its 8 iterations, many promising methodologies have emerged, but in particular, there has been an extensive acceptance of correlation filters based approaches with complex features and the deep convolutional neural networks due to the associated improvements in tracker performances [15], [16].

B. De-Interlacing

Interlacing can be described as a spatio-temporal subsampling technique, for TV broadcasts and video recording, that seeks to double the perceived frame rate without increasing the bandwidth. A frame in an interlaced video sequence typically consists of two fields captured consecutively. While one field consists of all odd-numbered lines of the frame, the second field contains all the even-numbered lines of the frame. The main benefit of interlacing stems from the fact that the human visual system is less sensitive to flickering details than to large-area flicker [17]. However, since two fields are captured at slightly different time intervals, video sequences consisting of fast moving objects exhibit tearing or interlacing artefacts, as shown in Fig. 1, which can be quite detrimental to the success of various computer vision tasks.

De-Interlacing algorithms, on the other hand, seek to interpolate two fields (half-frames) to recombine them into a corresponding full frame in order to alleviate the sub-sampling artefacts. This however is an ill-posed problem since two fields are captured at slightly different time intervals. De-interlacing algorithms can be classified into two categories:

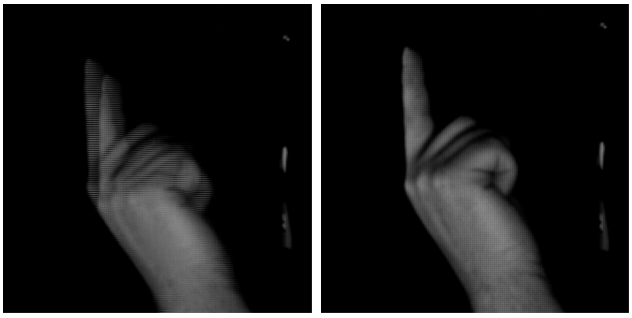


Fig. 1: An illustration of the tearing artifacts that arise due to interlacing (left) and their subsequent remediation by means of a de-interlacing algorithm (right).

techniques that aim to reproduce the whole frame from each of the odd-even fields independently and techniques that attempt to analyse the object motion before de-interlacing. While the former approach is associated with real-time performance but lower visual quality, the latter approach generally leads to better visual quality at a relatively higher computational cost.

C. Hand Movement Analysis Framework

The current work builds on a framework originally proposed in [18] and further developed in [9] with a view towards understanding the interactions of human subjects with stereoscopically rendered content on 3-D touch screens. Particularly, the authors were interested in deciphering user behaviour in instances when the stereoscopically rendered content floats in front of or behind the touch screen surface. The proposed hand movement analysis framework consisted of several blocks starting from setting up of the touch screen experiment and a collection of the video sequences. Video sequences were captured by means of a multi-camera setup involving a Sony HDRSR12 camera and a Mega Speed MS50K camera which captures videos at a frame rate of 25 fps (interlaced) and 500 fps (progressive) respectively. Based on the pointing actions carried out by 20 test subjects, paired normal and high speed video fragments were carefully extracted and pre-processed. Visual object trackers were then employed to reliably estimate the individual 2-D coordinates of the fingertips of the subjects for a particular pointing action from both the cameras across all the video frames. The obtained trajectory data were subsequently filtered and was used to compute the 3-D world coordinates of the fingertips with the aid of the intrinsic and extrinsic parameters of the cameras. The experimental setup for data collection as well as the subsequent hand movement analysis framework has been summarised in Fig. 2 [9].

Earlier studies on the framework were found wanting with regard to the tracker accuracy as the authors had to resort to unconventional strategies such as a backward tracking of the normal speed videos [5] which complicates the task of 3-D reconstruction. Additionally, in case of the high speed videos the movement between subsequent video frames could easily amount to less than a pixel and hence, the velocity and acceleration curves obtained from the trackers used in the

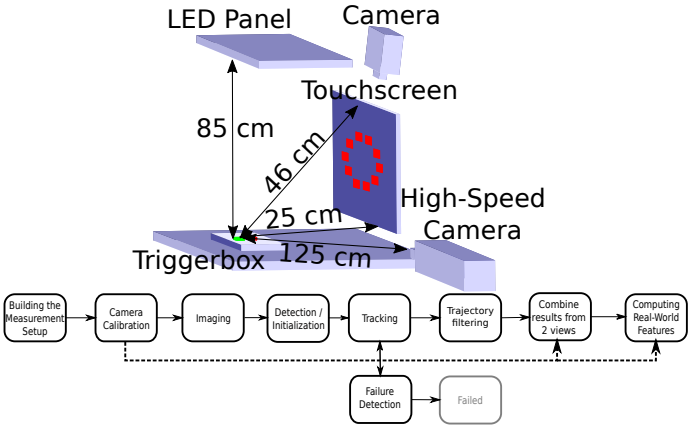


Fig. 2: The experimental setup (above) and the subsequent hand movement analysis framework (below)

earlier studies were also unstable and required a certain degree of a smoothing in order to obtain a reasonable approximation of the finger trajectory. While these limitations were addressed to a certain extent in one of the more recent iterations of published literature [19], without a highly accurate and sub-pixel precise estimate of the ground truth the authors could only then speculate about the superiority of such an approach. Furthermore, an unforeseen correlation was observed between the quality of the de-interlacing filter and the success of tracking in the case of the interlaced normal speed videos which warranted a broader study [19]. In the current work, we focus our attention primarily on these observations. Firstly, we propose a highly robust and accurate sub-pixel precise annotation scheme to establish ground truth fingertip locations for each of the video frames. Following which, we compare the performance of a modern deep learning based de-interlacing filter with the classical ffmpeg based de-interlacing filters with respect to the performance of the visual object trackers. Lastly, we briefly investigate the advantages of sub-pixel precise tracking over pixel precise tracking in terms of the observed trajectory, velocity and acceleration curves.

III. Experiments

A. Dataset Description

As described earlier, the video sequences considered in the present work emanate from two different cameras: A high speed camera that captures video sequences at 500 fps at a resolution of 800x601 and a normal speed RGB camera that records the user's pointing actions at 25 fps (interlaced) at a resolution of 1440x1080. While the initial HCI experiment captured pointing actions from 20 different human subjects we randomly sample 10 video sequences arising from different test subjects.

B. Sub-Pixel Precise Annotation

The incorporation of sub-pixel precise trackers causes traditional bounding box based annotation to be insufficient. Moreover, manual annotation frameworks such as LabelMe [20] add a degree of subjectivity to the task of annotation which could result in erroneous ground truth estimates, thereby corrupting any conclusions with regard to the tracking accuracy. Therefore, we devised an objective sub-pixel precise annotation scheme which employs a coarse-to-fine strategy for identifying the fingertip locations in any given frame.

At first, the pixel intensities are smoothed in the grayscale video frames by means of median filtering. This is necessary as the pixel intensities on the periphery of the fingertips were observed to be unstable. After this, the annotator draws a n-sided polygon by selecting n points around the ground truth fingertip position to mask out the region of interest. This is vital in order to prevent pixel intensities from the objects in the immediate vicinity of the finger such as the trigger box, or the touch screen to interfere with the soon-to-follow finger edge computation. After isolating the region of interest, a coarse estimate of the fingertip location is obtained based on the following assumption: the fingertip is the central uppermost nonzero pixel and the immediate vicinity of the fingertip consists of pixels with intensities zero. By means of the obtained coarse estimate of the fingertip, the left and right edge points and their corresponding midpoint are computed at a few discretized pixels (ranging from 1 to 10 pixels) below the coarse estimate. Then a line connecting the midpoint between the lowermost center point and the coarse fingertip estimate from earlier represents the longitudinal axis of the fingertip. The pixel intensities are then interpolated and smoothed along the fingertip axis and beyond followed by computing the gradient on the smoothed-interpolated pixel intensities along the axis. The point of minima, so observed from the computed intensity gradient, then corresponds to an accurate sub-pixel precise estimate of the fingertip position. The annotation strategy is summarised in Fig. 3.

C. De-interlacing and Tracking

The core objective of this study was to understand the effect of various de-interlacing filters on the accuracy of the trackers. While a study could be conducted on the reconstruction quality of various de-interlacing filters by applying them on the interlaced video sequences collected from the study, such a comparison however, would be incomplete without a ground truth reference progressive video. For this reason, the high speed video sequences captured at 500 frames per second were used instead. The high speed videos were converted into video sequences of 50 frames per second, which constituted as the ground truth and the resultant ground truth sequence was then interlaced for it to be de-interlaced by the various de-interlacing filters included in the study. A predetermined tracking algorithm was then employed with exactly the same initialization settings (bounding box estimate for the first frame, tracker parameters etc.) and was used to track the finger movements across both, the compiled video sequences as well as the generated ground truth progressive video. The tracker results were averaged over two runs on each of the video sequences.

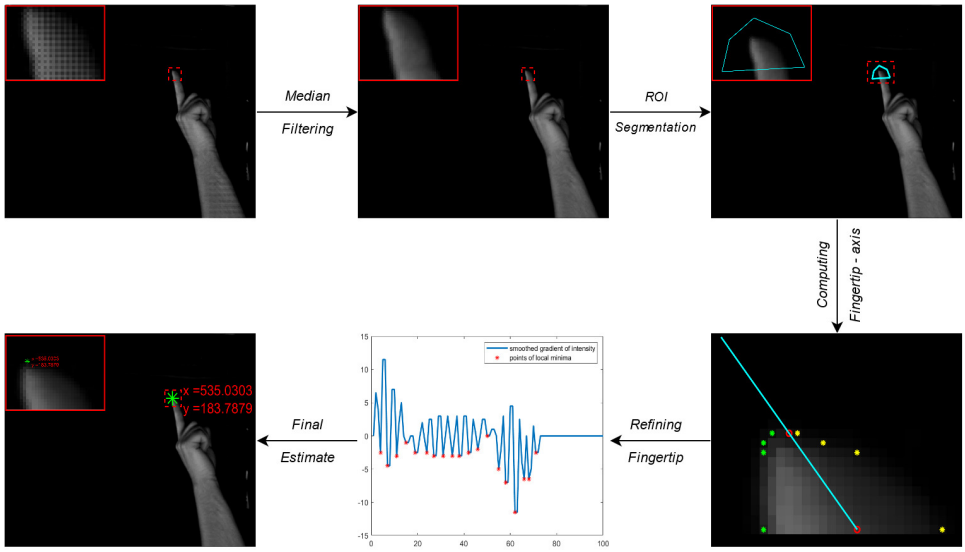


Fig. 3: A visual illustration of the proposed annotation framework.

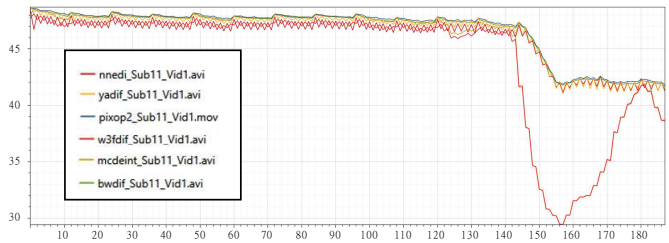
Accurate Tracking by Overlap Maximization (ATOM) [16] was selected as the tracking method since it outperformed the competing methods in an earlier study on similar data [19]. As for the de-interlacing algorithms, we selected five of the most popular classical de-interlacing algorithms available via FFmpeg [21] while also included a de-interlacing filter from a modern cloud based AI video enhancement tool, Pixop [10]. The ffmpeg based de-interlacers considered included the Weston 3 field de-interlacing filter (w3fdif), Motion Compensated De-Interlacing filter (mcdeint), Yet-another de-interlacing filter (yadif), Bob-Weaver de-interlacing filter (bwdif), and the Neural network edge directed interpolating (nnedi) filter.

1) Accurate Tracking by Overlap Maximization: The ATOM tracker aims to obtain a high-level understanding about the object's state in order to reliably track the target object [16]. This is achieved by segregating the task of tracking into two modules: a classification module that is trained online to predict a rough 2-D estimate for the target object and an estimation module that utilizes a pre-trained Intersection over Union (IoU) predictor network [22] and the rough estimates from the classification module to obtain a refined bounding box estimate for the target in the given frame. For more detailed description of the algorithm, see [16].

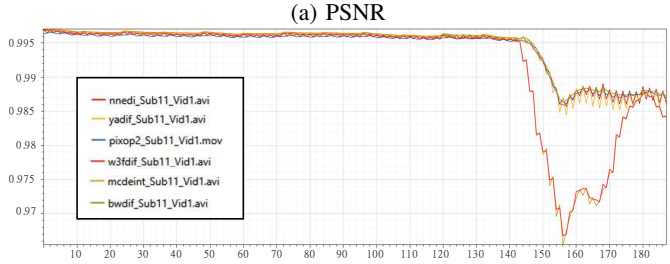
2) De-Interlacing Methodologies: The w3fdif was developed at the BBC R&D [23]. This filter employs the field dominance information in order to evaluate which of the odd or even fields to place first in the output. Yadif evaluates the pixels in the current, previous and next frame in order to recreate the field by means of edge directed interpolation. Two sets of outputs depending on its mode of operation can be obtained. If the mode is set as "send frame" the output consists of just one frame for each input frame and if the mode is set as "send field" the resulting output will consist of one frame for each of the fields and thereby, doubling the frame rate. The latter mode of operation was chosen based on a comparative study carried out earlier in the framework of this research, but the former mode was used in conjunction with the mcdeint de-interlacing as described in the ffmpeg documentation. The nnedi filter uses a predictor neural network along with neighbourhood pre-processing for obtaining the missing pixels by using information available only from the field being reconstructed. The bwdif filter employs a consolidation of the previously described de-interlacing filters such as yadif and w3fdif along with a few cubic interpolation algorithms for the de-interlacing of the interlaced video sequences. Lastly, the novel pixop tool [10] employs a deep convolutional neural network in order to optimally merge the effects of motion and temporal imperfections to generate high quality de-interlaced output.

D. Experimental Results

Central Location Error (CLE) was employed as the evaluation metric for the tracker performance. CLE is fundamentally a measure of the Euclidean distance between the estimated fingertip position of the tracker bounding box and the annotated fingertip position in any given frame. The adoption of CLE measure is justified in the context of the current study as it aligns well with the objective of tracking the fingertips and allows for comparability with the similar experiments carried out in the framework of touch screen usability experiment [5], [9], [19]. Additionally, average errors were computed along the x and y axis, respectively. The Tracker Success (TS) score was used to evaluate the success of the tracker across the entire video sequence. It corresponds to the percentage of video frames where the distance between the obtained fingertip estimate from the bounding box was less than a defined threshold from the ground truth fingertip position. The threshold was set as half the width and the height of the predicted bounding box window along the x and y axes respectively. Lastly, the average Peak Signal-to-Noise Ratio (PSNR) and the Structural

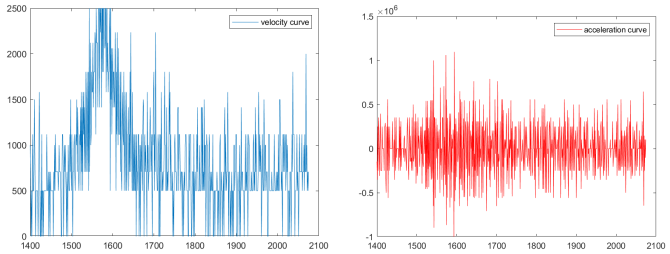


(a) PSNR

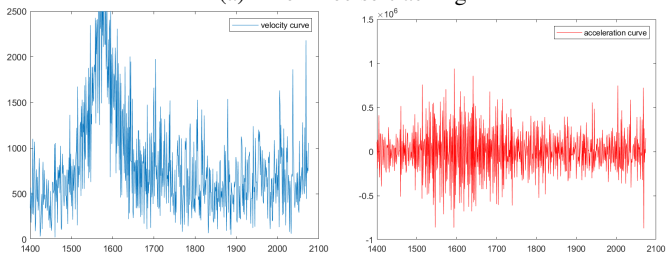


(b) SSIM

Fig. 4: Plots illustrating the quality of the video frames generated by various de-interlacing filters based on the PSNR and SSIM scores for one of the videos in our dataset. (Best viewed zoomed in at 250%)



(a) Pixel Precise tracking



(b) Sub-Pixel Precise tracking

Fig. 5: Illustration of the difference in the velocity and acceleration curves arising from difference in tracking precision on the High Speed Videos.

Similarity Index (SSIM) [24] of all the de-interlaced video frames were computed against the ground truth video frames in order to identify any existing correlation between the quality of de-interlacing and the tracker performance.

The results of the experiments averaged over the entire dataset are presented in Table I. The results demonstrate clear benefits of the deep learning based Pixop Video Enhancement Tool which provided the best performance in most categories including error along x, error along y, CLE, TS as well as PSNR. Fig. 4 further illustrates the de-interlacing capability of the various methodologies in terms of PSNR and SSIM, on a per frame basis, for one of the videos from our dataset. The drop in PSNR and SSIM scores in the latter stages of the video sequences can be explained by the fact that de-interlacing methods are a function of the amount of motion occurring in the frame and in the sample video considered here the fingertip only starts moving after the 140th frames.

E. Sub-Pixel Precision Trajectories

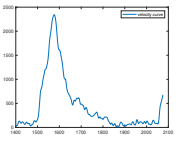
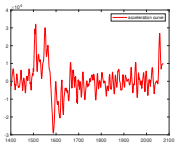
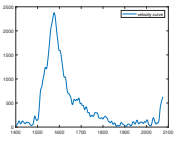
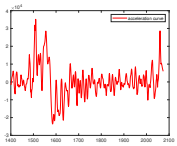
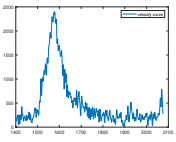
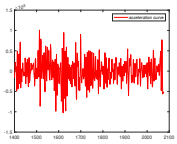
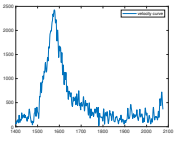
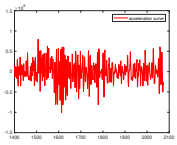
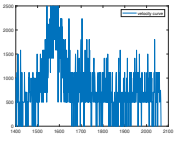
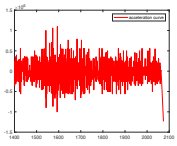
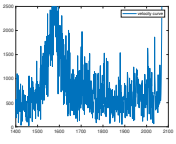
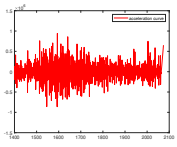
Prior studies involving the hand movement analysis framework [5], [9] relied exclusively on pixel precise bounding box estimates from the tracking algorithm. Such reliance implied that any information pertaining to the tracked fingertip's velocity and acceleration would be unreadable without prior smoothing of the raw trajectories. One of the reasons behind the unreadability of trajectories can be attributed to instances where the fingertip movement between consecutive frames is less than a pixel which would lead to rounding up of the bounding box estimates. We argue here that sub-pixel definition of the bounding box would be more pragmatic and would go some way towards remedying such errors due to rounding up. This advantage has been illustrated in Fig. 5 wherein one can observe the minor improvement in terms of the stability of the velocity and the corresponding acceleration curves.

The remaining inaccuracies in the fingertip location are due to the fact that the predicted bounding box in every frame is dependent on the confidence score of the IoU Net employed by the ATOM tracker which varies inevitably despite little to no movement of the finger in consecutive frames. Coupled with the fact that the fingertip is localised along the edge of the predicted bounding box rather than its center, it is clear to see that sub-pixel precision tracking on its own is insufficient for addressing all the errors. While trajectory smoothening approaches were employed in previous pixel precision studies, they have their own vulnerabilities. For instance, a large window size for smoothening could potentially imply a larger distortion to the raw true trajectories. The choice of window size is of vital importance, especially at the points of interest – the start and end of the finger movement in the given video sequences. In Table II, the effects of smoothening on the velocity and acceleration curves are compared and visualized for the high speed video sequences. It can be observed

TABLE I: Results of the experiment. The best performing de-interlacing filter has been highlighted in green while the tracking results on the Ground Truth video represents the upper limit of performance.

Videos	Δx (↓)	Δy (↓)	CLE (↓)	TS (↑)	PSNR (↑)	SSIM (↑)
Ground Truth	4.1267	1.6403	4.6823	97.1225 %	-	-
bwdif	6.3363	4.4126	8.1540	93.8732 %	44.4670	0.9920
mcdeint	34.5164	34.2731	50.6011	69.7259 %	37.4332	0.9882
nnedi	6.3381	3.7607	7.7691	94.0421 %	44.1282	0.9917
yadif	7.6301	4.2252	9.0007	93.3725 %	44.2774	0.9915
w3fdif	41.2231	38.3313	57.4718	69.4332 %	37.3993	0.9882
Pixop	4.9189	2.824	6.2703	95.9218 %	44.5841	0.9916

TABLE II: Comparing the effect of smoothing window size on the velocity and acceleration curves.

Nr.	Precision	Window Size	Velocity Curve	Acceleration Curve
1	Pixel	19		
2	Sub-Pixel	19		
3	Pixel	9		
4	Sub-Pixel	9		
5	Pixel	5		
6	Sub-Pixel	5		

that with a larger window size the curves arising from sub-pixel and pixel precise tracking tend to converge, but with a smaller window size the difference between them becomes pronounced. Smaller smoothing window sizes are preferable as they imply that in frames with high motion the deviation from the original trajectory is minimal.

IV. Conclusion

In this work, we explored a few intricate details concerning a computer vision based HCI experiment that were previously unexplored. Specifically, we address the issue of sub-pixel annotation while allowing for human errors via our semi-automatic annotation scheme. Apart from this, we also showcased the importance of the choice of de-interlacing algorithms on the tracker accuracy. Lastly, we argued the relevance of sub-pixel precision in the context of our HCI experiment while also addressing the shortcomings and the reason behind them. While we based our findings on a very specific HCI experiment, our findings can be easily interpolated to other computer vision based HCI experiments as well as practical machine vision tasks relying on interlaced scan sensors and precision object tracking.

References

[1] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 52–73, 2007.

[2] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[3] D. Valkov, A. Giesler, and K. Hinrichs, "Evaluation of depth perception for touch interaction with stereoscopic rendered objects," in *International Conference on Interactive Tabletops and Surfaces*, 2012, pp. 21–30.

[4] Y. Fang, W. Kang, Q. Wu, and L. Tang, "A novel video-based system for in-air signature verification," *Computers & Electrical Engineering*, vol. 57, pp. 1–14, 2017.

[5] V. Lyubanenko, T. Kuronen, T. Eerola, L. Lensu, H. Kälviäinen, and J. Häkkinen, "Multi-camera finger tracking and 3d trajectory reconstruction for hci studies," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2017, pp. 63–74.

[6] L.-W. Chan, H.-S. Kao, M. Y. Chen, M.-S. Lee, J. Hsu, and Y.-P. Hung, "Touching the void: Direct-touch interaction for intangible displays," in *Conference on Human Factors in Computing Systems*, 2010, pp. 2625–2634.

[7] T. Kuronen, "Moving object analysis and trajectory processing with applications in human-computer interaction and chemical processes," Ph.D. dissertation, Lappeenranta University of Technology, 2018.

[8] "Interlaced cameras: a renaissance?" https://www.stemmer-imaging.com/media/uploads/websites/documents/tech-tips/en-Tech-Tip-Interlacing-TTCAM2-201108.pdf, accessed: 2021-05-01.

[9] T. Kuronen, T. Eerola, L. Lensu, J. Häkkinen, and H. Kälviäinen, "3d hand movement measurement framework for studying human-computer interaction," in *International Conference Cyber-Physical Systems and Control*. Springer, 2019, pp. 513–524.

[10] "Pixop: AI video enhancement and upscaling in the cloud," https://www.pixop.com/, accessed: 2021-05-30.

[11] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2411–2418.

[12] Q. Wang, F. Chen, W. Xu, and M.-H. Yang, "An experimental comparison of online object-tracking algorithms," in *Wavelets and Sparsity XIV*, vol. 8138. International Society for Optics and Photonics, 2011, p. 81381A.

[13] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognition*, vol. 76, pp. 323–338, 2018.

[14] "VOT Challenge Homepage," https://www.votchallenge.net/, accessed: 2021-05-25.

[15] T. Böttger, M. Ulrich, and C. Steger, "Subpixel-precise tracking of rigid objects in real-time," in *Scandinavian Conference on Image Analysis*. Springer, 2017, pp. 54–65.

[16] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Atom: Accurate tracking by overlap maximization," in *Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4660–4669.

[17] E. Engstrom, "A study of television image characteristics: Part two: Determination of frame frequency for television in terms of flicker characteristics," *Institute of Radio Engineers*, vol. 23, no. 4, pp. 295–310, 1935.

[18] T. Kuronen, T. Eerola, L. Lensu, J. Takatalo, J. Häkkinen, and H. Kälviäinen, "High-speed hand tracking for studying human-computer interaction," in *Scandinavian Conference on Image Analysis*. Springer, 2015, pp. 130–141.

[19] A. Nawaz, "Hand tracking with sub-pixel precision," Master's thesis, Lappeenranta University of Technology, 2020.

[20] K. Wada, "labelme: Image Polygonal Annotation with Python," https://github.com/wkentaro/labelme, accessed: 2021-07-21.

[21] "FFmpeg," https://ffmpeg.org/, 2000, accessed: 2021-05-30.

[22] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 784–799.

[23] M. Weston, "Interpolating lines of video signals," Dec. 6 1988, uS Patent 4,789,893.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.