Deep SORT with Non-linear State Estimation: Exploring Kalman Filters

Abhi Shukul
Institution1
Institution1 address
firstauthor@i1.org

Yamato Miura
Institution2
First line of institution2 address
secondauthor@i2.org

Abstract

The original implementation of DeepSORT utilizes a standard Kalman filter. In this project, we consider alternative types of Kalman Filters within the DeepSORT (Deep Simple Online and Realtime Tracking) framework to track sports balls, particularly in scenarios where ball trajectories exhibit non-linear behavior. The primary problem was to accurately track the movement of sports balls in sports videos where traditional linear models might not suffice due to the nature of ball dynamics (like spin, air resistance, etc.) or occlusions.

The idea stemmed from the need to improve object tracking, especially in sports analytics, where precise trajectory data is crucial for performance analysis, strategy optimization, and player evaluation. DeepSORT, with non-linear Kalman filters, offers a promising solution for robust object tracking, even in complex scenarios.

1. Introduction

The goal of object tracking is to be able to detect an object (e.g. a ping pong ball) and track its motion in a video by drawing bounding boxes around it. A prevalent approach for solving this task is the Simple Online and Realtime Tracking algorithm, an approach that utilizes Kalman filters. This approach can be further refined using deep learning for object identification and incorporating YOLO (You Only See Once) object detection.

While DeepSORT with YOLO can be fairly effective for general object tracking–such as tracking cars driving on a highway, its use of the standard Kalman filter means that it tends to work better for objects that have a linear motion. This is because the standard Kalman filter assumes the the motion of the object is linear. The goal of our projeect was to investigate alternative non-linear Kalman filters such as the extended Kalman filter and unscented Kalman filter and try to incorporate them into DeepSORT to see if they can improve performance for non-linear object tracking. In order to test the different Kalman filters, we focused on tracking ping-pong balls since they tend to exhibit mon-linear motion.

2. Background

A prevalent approach for solving the task of tracking sports balls is the Simple Online and Realtime Tracking (SORT) algorithm, as outlined in https://arxiv.org/abs/1602.00763 **CHANGE TO REFERENCE**. Furthermore, in previous research in computer vision explored various approaches to object tracking, ranging from optical flows and Kalman Filters as SORT does. Furthermore, this approach can be improved using a deep neural network for object identification, as described in https://arxiv.org/pdf/1703.07402 **CHANGE TO REFERENCE**. DeepSORT is a technique that combines both methods, where the object is classified with Convolutional Neural Networks and the object is tracked using a SORT algotithm, which utilizes Kalman Filters. Kalman Filters have been widely used for state estimation in linear systems. https://arxiv.org/abs/1703.07402 **CHANGE TO REFERENCE**. We hope to provide a solution to track ping pong balls effectively. However, we can assume the ping pong ball can be modeled as a linear system if the ball moves in a linear trajectory with relatively low acceleration. Trajectories that are relatively parabolic, curved due to ball dynamics, or trajectories where the ball is occluded briefly can cause the linear system to terribly mispredict the next state of the ball. Thus, we aim to improve the state estimation in DeepSORT using various Kalman Filters for non-linear state estimation.

3. Methodology

In order to track objects across frames, we need to be able to have our software examine a frame and detect the object we want to track. To detect objects, we utilized YOLO (You Only Look Once), a real-time object detection algorithm that is adept at performing detection in videos with low latency and high accuracy. YOLO, developed by Ultralytics, comes pre-trained on the COCO dataset, a general-

purpose dataset containing a variety of classes such as animals, vehicles, and food. It performs poorly on things like sports balls/ping pong balls, which it has not been explicitly trained on; when we ran our test video using the pre-trained version of YOLO, it detected cars and motorcycles (which were not actually there) in the video rather than ping pong balls. To improve performance, we trained the YOLO model for 50 epochs on a ping pong dataset. Additionally, the pre-trained YOLO model is not adept at object tracking, as it performs detection frame-by-frame and does not cache detection information across frames. Tracking across frames is a key part of sports ball tracking; in order to incorporate this, we used the DeepSORT framework.

DeepSORT is a framework that utilizes an object classification framework using CNNs and tracks the object using Kalman Filters [1]. DeepSORT utilizes deep learning in addition to CNNs to accurately track fast-moving objects - a place where YOLO and DeepSORT predecessors like SORT fall short. The original Kalman Filter utilized in the paper and the YOLO model is meant for objects that move in a relatively linear fashion. However, this assumption is not true for sports balls and does not account for ball dynamics, such as when spin is induced by how sports players interact with the balls. Initially, before we introduced the new Kalman filters, the model struggled to track the ball once it changed direction, or when it was hit quickly. Thus, we have modified the Kalman Filters utilized in DeepSORT and benchmarked the performance of our implementation's ability to track sports balls that perform state estimation for nonlinear systems, such as the case where the ball moves in a non-linear trajectory.

We found an implementation of YOLOv8 + DeepSORT [1], which uses the standard Kalman filter, and used it as a baseline. After training YOLO on the Ping-Pong dataset [2] for 50 epochs and a learning rate of 0.001, we applied 3 different versions of DeepSORT, each using a different Kalman Filter (the pre-existing linear filter, and 2 non-linear versions) and compared the results we got from each model.

We improved the YOLO model in a few different ways:

3.1. Data Pre-Processing

We added image sharpening, wherein each frame of our test video was sharpened using a 3x3 filter and cross-correlation.

3.2. Non-linear Kalman Filters

We attempted implementations of two different non-linear versions of Kalman filters and integrated of it within the existing code.

The Extended Kalman Filter (EKF) is the simplest of the three filters. It involves attempting to linearize any non-linearities in the state model by taking the derivative and making a linear approximation. EKF works well in cases where nonlinearities in the system are subtle but does not work as effectively for more egregious nonlinearities.

The unscented Kalman Filter (UKF) works better for more egregious nonlinearities. They use a type of transformation called an "unscented transformation" as part of their algorithm to estimate bounding boxes. The algorithm outputs "sigma points", which are used to gather the mean and the covariance of the different state parameters and which are then used to better estimate bounding boxes on an object that is moving nonlinearly.

We looked into pre-existing online guides and relevant literature

We faced significant challenges in fully integrating these improvements within the pre-existing framework within the given timeframe, more details are discussed in

3.3. Motion Filtering

A major issue we ran into was that the model was identifying any small, circular, white object as a ping pong ball, even if it was just part of the background of the video. For example, it detected portions of In order to improve this, we implemented motion filtering. We added to DeepSORT the ability to cache previous positions of objects and use that to calculate an approximation for the speed of the objects being tracked. We then filtered the bounding boxes by the speed of the object they were tracking, and only drew them if the object was moving above a particular speed. We tuned the threshold for the speed, running it multiple times until we found a value that blocked out as much background as we could while not impacting the detection of the actual ball.

Clearly, the box loss decreased as a result of our training. We started from a loss of approximately 2.00, and ended with a final loss of 1.245.

Qualitatively, when we used the inbuilt weights (from training using the COCO dataset) on our test video, we observed that the model was not able to detect any balls, and instead was detecting random objects in the background as cars (see left image below; the ball is circled in red and the purple bounding boxes are what was actually tracked). After training, using the base Kalman filters already provided in the repository we cloned, the model was able to track the ball much better (see right image below), although it still failed to track when the ball was moving in a nonlinear fashion (right after the paddle hits the ball).

3.4. Dual submission

Please refer to the author guidelines on the CVPR 2022 web page for a discussion of the policy on dual submissions.

[1] https://github.com/MuhammadMoinFaisal/YOLOv8-DeepSORT-Object-Tracking

[2] https://universe.roboflow.com/face-imsn4/ping-pong-ball-qsdtm/dataset/8/images/426479f265c7c5bd0143df01f669febc

(a) label 1

(b) label 2

Figure 1. result from applying deepSort with base Kalman filter to a Ping Pong video. Observe from the image on the left that the tracker fails shortly after the ball comes into contact with the paddle (because of non-linear motion)

3.5. Paper length

Papers, excluding the references section, must be no longer than eight pages in length. The references section will not be included in the page count, and there is no limit on the length of the references section. For example, a paper of eight pages with two pages of references would have a total length of 10 pages. **There will be no extra page charges for CVPR 2022.**

Overlength papers will simply not be reviewed. This includes papers where the margins and formatting are deemed to have been significantly altered from those laid down by this style guide. Note that this LaTeX guide already sets figure captions and references in a smaller font. The reason such papers will not be reviewed is that there is no provision for supervised revisions of manuscripts. The reviewing process cannot determine the suitability of the paper for presentation in eight pages if it is reviewed in eleven.

3.6. The ruler

The LaTeX style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document using a non-LaTeX document preparation system, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera-ready copy should not contain a ruler. (LaTeX users may use options of cvpr.sty to switch between different versions.)

Reviewers: note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (*e.g.*, this line is 087.5), although in most cases one would expect that the approximate location will be adequate.

3.7. Paper ID

Make sure that the Paper ID from the submission system is visible in the version submitted for review (replacing the "*****" you see in this document). If you are using the LaTeX template, **make sure to update paper ID in the appropriate place in the tex file**.

3.8. Mathematics

Please number all of your sections and displayed equations as in these examples:

$$E = m \cdot c^2 \tag{1}$$

and

$$v = a \cdot t. \tag{2}$$

It is important for readers to be able to refer to any particular equation. Just because you did not refer to it in the text does not mean some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like "the equation second from the top of page 3 column 1". (Note that the ruler will not be present in the final copy, so is not an alternative to equation numbers). All authors will benefit from reading Mermin's description of how to write mathematics: http://www.pamitc.org/documents/mermin.pdf.

3.9. Blind review

Many authors misunderstand the concept of anonymizing for blind review. Blind review does not mean that one must remove citations to one's own work—in fact it is often impossible to review a paper unless the previous citations are known and available.

Blind review means that you do not use the words "my" or "our" when citing previous work. That is all. (But see below for tech reports.)

Saying "this builds on the work of Lucy Smith [1]" does not say that you are Lucy Smith; it says that you are building on her work. If you are Smith and Jones, do not say "as we show in [7]", say "as Smith and Jones show in [7]" and at the end of the paper, include reference 7 as you would any other cited work.

An example of a bad paper just asking to be rejected:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of our previous paper [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Removed for blind review

An example of an acceptable paper:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of the paper of Smith *et al*. [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Smith, L and Jones, C. "The frobnicatable foo filter, a fundamental contribution to human knowledge". Nature 381(12), 1-213.

If you are making a submission to another conference at the same time, which covers similar or overlapping material, you may need to refer to that submission in order to explain the differences, just as you would if you had previously published related work. In such cases, include the anonymized parallel submission [5] as supplemental material and cite it as

> [1] Authors. "The frobnicatable foo filter", F&G 2014 Submission ID 324, Supplied as supplemental material `fg324.pdf`.

Finally, you may feel you need to tell the reader that more details can be found elsewhere, and refer them to a technical report. For conference submissions, the paper must stand on its own, and not *require* the reviewer to go to a tech report for further details. Thus, you may say in the body of the paper "further details may be found in [6]". Then submit the tech report as supplemental material. Again, you may not assume the reviewers will read this material.

Sometimes your paper is about a problem which you tested using a tool that is widely known to be restricted to a single institution. For example, let's say it's 1969, you have solved a key problem on the Apollo lander, and you believe that the CVPR70 audience would like to hear about your solution. The work is a development of your celebrated 1968 paper entitled "Zero-g frobnication: How being the only people in the world with access to the Apollo lander source code makes us a wow at parties", by Zeus *et al*.

You can handle this paper like any other. Do not write "We show how to improve our previous work [Anonymous, 1968]. This time we tested the algorithm on a lunar lander [name of lander removed for blind review]". That would be silly, and would immediately identify the authors. Instead write the following:

> We describe a system for zero-g frobnication. This system is new because it handles the following cases: A, B. Previous systems [Zeus et al. 1968] did not handle case B properly. Ours handles it by including a foo term in the bar integral.
>
> ...
>
> The proposed system was integrated with the Apollo lunar lander, and went all the way to the moon, don't you know. It displayed the following behaviours, which show how well we solved cases A and B: ...

As you can see, the above text follows standard scientific convention, reads better than the first version, and does not explicitly name you as the authors. A reviewer might think it likely that the new paper was written by Zeus *et al*., but cannot make any decision based on that guess. He or she would have to be sure that no other authors could have been contracted to solve problem B.

FAQ

Q: Are acknowledgements OK?
A: No. Leave them for the final copy.

Q: How do I cite my results reported in open challenges?
A: To conform with the double-blind review policy, you can report results of other challenge participants together with your results in your paper. For your results, however, you should not identify yourself and should not mention your participation in the challenge. Instead present your results referring to the method proposed in your paper and draw conclusions based on the experimental comparison to other results.

3.10. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	conf_a

See The TEXbook, p165.

The space after *e.g*., meaning "for example", should not be a sentence-ending space. So *e.g*. is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

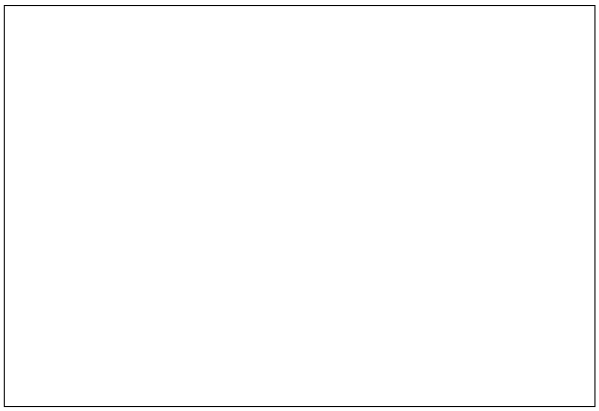


Figure 2. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al.*" (not "*et. al.*" as "*et*" is a complete word). If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al*. However, use it only when there are three or more authors. Thus, the following is correct: "Frobnication has been trendy lately. It was introduced by Alpher [1], and subsequently developed by Alpher and Fotheringham-Smythe [2], and Alpher *et al*. [3]."

This is incorrect: "... subsequently developed by Alpher *et al*. [2] ..." because reference [2] has just two authors.

4. Formatting your paper

All text must be in a two-column format. The total allowable size of the text area is $6\frac{7}{8}$ inches (17.46 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be $1\frac{1}{8}$ inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately $1\frac{5}{8}$ inches (4.13 cm) from the bottom edge of the page.

4.1. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area $6\frac{7}{8}$ inches (17.46 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Page numbers should be in the footer, centered and $\frac{3}{4}$ inches from the bottom of the page. The review version should have page numbers, yet the final version submitted as camera ready should not show any page numbers. The LaTeX template takes care of this when used properly.

4.2. Type style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title $1\frac{3}{8}$ inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. $\frac{1}{6}$ inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figs. 2 and 3. Short captions should be centred.
Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

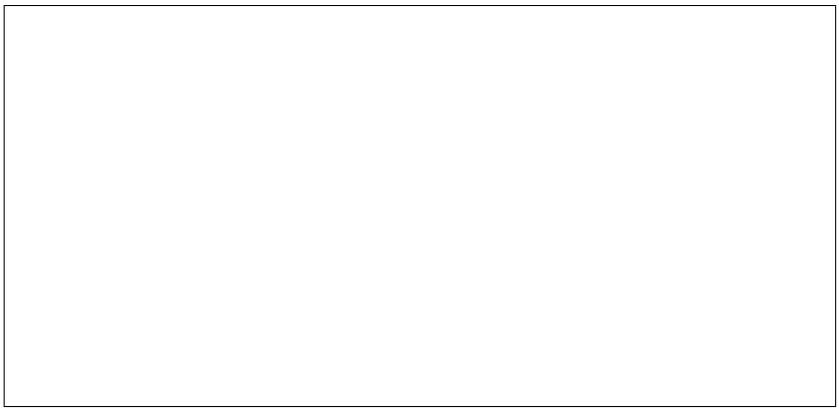
4.3. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

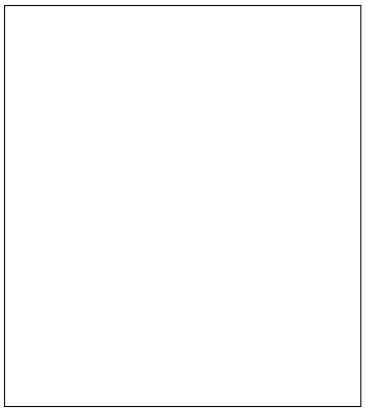
4.4. Cross-references

For the benefit of author(s) and readers, please use the

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.



(a) An example of a subfigure.



(b) Another example of a subfigure.

Figure 3. Example of a short caption, which should be centered.

```
\cref{...}
```

command for cross-referencing to figures, tables, equations, or sections. This will automatically insert the appropriate label alongside the cross-reference as in this example:

> To see how our method outperforms previous work, please see Fig. 2 and Tab. 1. It is also possible to refer to multiple targets as once, *e.g*. to Figs. 2 and 3a. You may also return to Sec. 4 or look at Eq. (2).

If you do not wish to abbreviate the label, for example at the beginning of the sentence, you can use the

```
\Cref{...}
```

command. Here is an example:

> Figure 2 is also quite important.

4.5. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [5]. Where appropriate, include page numbers and the name(s) of editors of referenced books. When you cite multiple papers at once, please make sure that you cite them in numerical order like this [1, 2, 4–6]. If you use the template as advised, this will be taken care of automatically.

4.6. Illustrations, graphs, and photographs

All graphics should be centered. In LaTeX, avoid using the `center` environment for this purpose, as this adds potentially unwanted whitespace. Instead use

```
\centering
```

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Results. Ours is better.

at the beginning of your figure. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths that render effectively in print. Readers (and reviewers), even of an electronic copy, may choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.pdf}
```

4.7. Color

Please refer to the author guidelines on the CVPR 2022 web page for a discussion of the use of color in your document.

If you use color in your plots, please keep in mind that a significant subset of reviewers and readers may have a color vision deficiency; red-green blindness is the most frequent kind. Hence avoid relying only on color as the discriminative feature in plots (such as red *vs*. green lines), but add a second discriminative feature to ease disambiguation.

5. Final copy

You must include your signed IEEE copyright release form when you submit your finished paper. We MUST have this form before your paper can be published in the proceedings.

Please direct any questions to the production editor in charge of these proceedings at the IEEE Computer Society Press: https://www.computer.org/about/contact.

References

[1] FirstName Alpher. Frobnication. *IEEE TPAMI*, 12(1):234–778, 2002. 5, 6

[2] FirstName Alpher and FirstName Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003. 5, 6

[3] FirstName Alpher, FirstName Fotheringham-Smythe, and FirstName Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004. 5

[4] FirstName Alpher and FirstName Gamow. Can a computer frobnicate? In *CVPR*, pages 234–778, 2005. 6

[5] FirstName LastName. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as supplemental material `fg324.pdf`. 4, 6

[6] FirstName LastName. Frobnication tutorial, 2014. Supplied as supplemental material `tr.pdf`. 4, 6