Quantitative and Qualitative Evaluation of Visual Tracking Algorithms using Statistical Tests

Michael Pound, Asad Naeem, Andrew French & Tony Pridmore
School of Computer Science, The University of Nottingham
{mpp,azn,apf,tpp}@cs.nott.ac.uk

Abstract

Performance evaluation of visual tracking algorithms is a complex task requiring consideration of the robustness, accuracy and failure modes of any proposed method. Both artificial and real data sets are typically employed, with quantitative, distance-based measures of tracking accuracy supported by qualitative, manual analysis of robustness and failure modes. Failure is usually taken to mean dissociation of the tracker with its target, and is identified by eye. Although distance measures are valid given artificial data, manually generated ground truth is not sufficient to allow them to be used reliably on real data sets. This paper presents an alternative evaluation methodology in which quantitative measures, in the form of statistical tests, are applied to the evaluation of both accuracy and robustness. These tests then form the basis of a tool that automatically identifies tracking failures, focusing attention on a well-defined set of events.

1. Introduction

Object tracking is both an important visual skill in its own right and an integral part of higher-level tasks such as event detection, surveillance, etc. A large number of algorithms have been proposed for tracking objects in image sequences. Each has its own strengths and weaknesses and, as yet, no single algorithm or approach has been shown to be generally applicable. Visual tracking remains an active research topic.

In this context, as elsewhere in computer vision, performance evaluation is crucial. The evaluation of visual tracking algorithms is a complex task requiring consideration of the accuracy, robustness and failure modes of any proposed method. Both absolute performance values over those test sets and comparisons between competing trackers are needed.

Tracking accuracy is generally assessed by measuring the distance between the algorithm’s estimated values of the tracked variables and appropriate ground truth data. Ground truth may be acquired by generating artificial test data, so that true values are known a priori, or by marking up real image sequences, a time-consuming and laborious task. Root mean squared error (RMS) is perhaps the most commonly used distance measure [1-4], providing an absolute measure of accuracy that can be compared with other values achieved over the same data set.

Robustness and failure modes are typically evaluated only qualitatively. The proposed algorithm is applied to a set of test sequences, usually combining real and artificial data, and the number and nature of its failures are discussed. Failure is usually taken to mean dissociation of the tracker with its target, and is identified by eye.

Many factors affect the performance of tracking algorithms, each in a variety of ways depending upon the details of the tracker. Lighting changes, the presence of clutter, the motion of the tracked object, camera shifts, target speed, partial and complete occlusions etc. should all be considered. Recent studies have proposed the use of standard test sets designed to sample these effects. In what follows, we assume that suitable test data is available and focus on the analysis and interpretation of the resulting ground truth and tracker output.

The level of control present in the generation of artificial data allows distance measures to be used to reliably assess tracking accuracy. Artificial data, however, can only approximate real world tracking problems and so cannot support a complete assessment. Real image sequences must be included in any comprehensive evaluation protocol. The evaluation of tracking algorithms using real image sequences is problematic in a number of ways.

First, although distance measures are valid given artificial data, the available ground truth is not generally sufficient to allow them to be used reliably on real data sets. Real test sequences are typically ground-truthed manually, by pausing at each frame while a user points out where the object is in that
frame. This process is subjective and prone to error – even if a well-defined image feature is used as a target the user is unlikely to mark it accurately throughout the large number of frames making up a typical test sequence. Care must also be taken to avoid quantisation error; manual ground-truthing typically provides estimates to the nearest pixel, while tracking algorithms provide real-valued estimates of position and other object parameters. In some circumstances a single-frame feature detection algorithm may be used to recover object location. While this can reduce workload, the resulting ground truth can still be expected to contain positional errors dependent on the algorithm employed.

A higher degree of subjectivity can be expected during the evaluation of robustness and failure modes, where both the identification and interpretation of key events is carried out by the assessor. Though more weight is usually given to the assessment of robustness and failure using real data, experiments performed on artificial data are equally subjective.

This paper presents an alternative evaluation methodology in which quantitative measures, in the form of statistical tests, are applied to the evaluation of accuracy, robustness and failure modes of visual trackers. Though we illustrate the proposed techniques using real data image sequences, they are equally applicable to artificial test data.

Issues in the assessment of robustness are addressed in Section 2, where it is shown that McNemar’s test can be used to provide a statistical comparison of the robustness of competing algorithms. Section 3 considers the evaluation of tracking accuracy given real image sequences and their associated, potentially errorful ground truth. Statistical tests, specifically permutation tests, are again proposed as a route to reliable quantitative evaluation. These tests form the basis of a computational tool, presented in Section 4, which supports the qualitative investigation of failure modes, focusing attention on a well-defined set of events. Finally, conclusions are drawn in Section 5.

2. Robustness

The robustness of a visual tracker is a measure of the extent to which it remains associated with the (correct) target throughout the test sequences. At present all trackers can reasonably be expected to lose their target at some point, the purpose of any performance evaluation scheme is to investigate the frequency with which and the conditions under which this occurs. A typical robustness test involves applying the algorithm to a set of image sequences, noting the points at which it fails and discussing why each failure, or a representative set of such failures, occurred. Comparative analysis is achieved by pointing out situations in which tracker A lost the target while tracker B maintained its lock, and vice versa. Robustness analysis is therefore qualitative, and largely subjective. Statistical tests exist, however, which can provide principled, quantitative measures of the relative robustness of tracking algorithms.

We propose the use of McNemar’s test [5] to provide quantitative, comparative analysis of the robustness tracking algorithms. McNemar’s statistic is a form of chi-square test for matched paired data. Let \( N_{sf} \) and \( N_{fs} \) denote failure and success respectively. McNemar’s statistic is then:

\[
X^2 = \frac{(|N_{sf} - N_{fs}| - 1)^2}{N_{sf} + N_{fs}}
\]

The Z score (standard score) is obtained as:

\[
Z = \frac{|N_{sf} - N_{fs}| - 1}{\sqrt{N_{sf} + N_{fs}}}
\]

If the two algorithms give similar results then \( Z \) will tend to zero (though \( Z \) tends to infinity if the two algorithms perform exactly the same). As their results diverge, \( Z \) increases. Confidence limits can be associated with the \( Z \) value [5].

To apply McNemar’s, a definition of success and failure is required. At present, any tracker may reasonably be expected to fail at some point. We therefore consider tracker A to have succeeded and tracker B to have failed if algorithm A maintains tracking for a greater proportion of a given image sequence, measured from the beginning of the sequence and using the same starting parameters. In effect we define success to be tracking as long as the more successful of the two trackers, and employ test sequences that are long enough and complex enough to force failure. In our applications of McNemar’s to date [1,6] loss of target is identified by eye, but clearly defined as there being no association between any part of the target and any aspect of the tracker. Future robustness tests will employ the tool described in section 4 to identify failure.

McNemar’s statistic is computed over a representative, and ideally standard and publically
available set of image sequences, requiring only 30 sequences to produce a statistically reliable result.

McNemar’s test was recently [6] applied to a set of 36 assorted image sequences (available from http://www.cs.nott.ac.uk/~azn/kams_bm vc.htm) to provide quantitative comparison of a novel tracking algorithm, KAMS, which combines annealed particle filtering [7] with the Kernel Mean Shift algorithm [3]. KAMS was compared with pure Kernel Mean Shift [3], Condensation [8] and Annealed particle filters [7], and previous hybrid algorithms due to Maggio and Cavallaro [9] (known as “Hybrid”) and Naeem et al [1] (known as “SOK”). With Z scores shown in Table 1, KAMS was found significantly more robust than each of the five reference algorithms, over the test set, with a confidence of 99.5%.

Table 1: Z score comparisons of KAMS with the other five algorithms.

<table>
<thead>
<tr>
<th></th>
<th>Kams vs. Condensation</th>
<th>Kams vs. Mean Shift</th>
<th>Kams vs. Annealing</th>
<th>Kams vs. Hybrid</th>
<th>Kams vs. SOK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>2.9104</td>
<td>5.1265</td>
<td>3.8013</td>
<td>4.8280</td>
<td>3.5907</td>
</tr>
</tbody>
</table>

Detailed qualitative analysis of the causes of tracking failure will always be required to identify the strengths and weaknesses of tracking algorithms and to identify directions for future research. We believe, however, that statistical tests are valuable in the formal assessment of performance, and that McNemar’s statistic, used as described here, is a valuable tool in the comparative analysis of tracking algorithms.

In addition to supporting a formal comparison, N
f and N
f provide further useful information: if these values are both large, then algorithm A tends to succeed where algorithm B fails and vice versa. This is useful to know, as it indicates the presence of complimentary approaches and so a potential for improved tracking through their combination.

3. Accuracy

Previous publications on tracking have predominantly used distance measures as a means for quantitative evaluation of tracking accuracy. For artificial data where the ground truth is known, distance can be an invaluable tool. With a user defined ground truth, however, errors in ground truth can cause the distance measure to become unreliable. The weaknesses of distance measures in determining tracker accuracy can be overcome by instead using statistical measures - the output of a successful tracker will be significantly similar and close to the ground truth data. Permutation tests can be used to determine whether or not this is the case.

It is common when analysing data to assume a null hypothesis about that data, such as assuming there is no difference between two data sets. Such a hypothesis is assumed true until sufficient evidence is provided that it is likely to be false, and can therefore be rejected. The null hypothesis in this case is that the data from the ground truth and tracker output come from the same distribution. If this is the case, it can be said that both data sets are similar, i.e. the tracker has been successful with respect to the ground truth. Rejecting the null hypothesis would suggest that the data sets are significantly different, i.e. that the tracked data does not match the ground truth path sufficiently well.

A common statistical test for validating or refuting the null hypothesis is a permutation test. This requires the assumption that the data from both sets are exchangeable under the null hypothesis. Details of the test can be seen in Figure 1.

**Permutation Test**

1. Data is taken from two sets A and B of size n
A and n
B respectively.
2. Calculate the observed test statistic T(obs).
3. Permutations. Calculate the distribution of the test statistic T under many rearrangements of the labels on the observed data points. In detail:
   a. Pool observations of groups A and B.
   b. Sample from these pooled values, n
A observations without replacement. The remaining observations form group n
B.
   c. Record the test statistic under this rearrangement.
   d. Repeat for all possible rearrangements or as many at random as are computationally efficient.
4. The two-tailed p-value is calculated as the proportion of sampled permutations where the test statistic was greater than or equal to abs(T(obs)).

**Figure 1. Overview of the Permutation test algorithm**

It is possible to use a variety of test statistics in a permutation test. One should be chosen that produces...
evidence against the null hypothesis. In the current context we are primarily interested in comparing two paths – the ground truth and that taken by a proposed tracking algorithm. As a result, we focus on test statistics that examine whether or not these paths are the same shape and in the same place (on the image plane).

If the tracker output and ground truth are significantly correlated, the tracker can be said to be determining the movement of the object correctly. Spearman’s rank correlation is a special case of the Pearson product-moment correlation coefficient, where data is first converted into ranks. This removes the assumption that relationship between the two datasets is linear. In tracker evaluation, negative correlations can be disregarded, as they arise when the tracker output produces a result similar to the reverse of the ground truth. A positive correlation must only be accepted if the value is significantly close to 1. A permutation test with Spearman rank correlation as a test statistic can distinguish a significantly high correlation. During the test, the data is rearranged many times, at each rearrangement the correlation is recorded. A p-value is then calculated from the distribution. This value is compared to a chosen significance level, a significant correlation shows that the tracker output is strongly correlated to the ground truth – i.e. the two paths are the same shape.

Spearman rank correlation can be applied to a wide variety of data types, in our experiments to date we have performed correlations only on target position estimates made within the image plane. Tracker estimates of x-coordinate are correlated with ground truth x values, and tracked y-coordinate with ground truth y values.

The weakness of Spearman rank correlation is its inability to respond to the distance between two data sets. The method returns a high correlation from two object paths that are the same shape but widely separated in the image plane. In some circumstances this may be an advantage, but in general tracker evaluation it is not. A second permutation test based on the mean values of these sets can rectify this. In the second test, the statistic used is the difference between the mean values of groups \( nA \) and \( nB \) - the tracking data and the ground truth. A significantly small distance between these two means provides further evidence that the null hypothesis is correct by requiring the two paths to be close to each other.

The combination of these two permutation tests produces a powerful comparison of two data sets. They provide us with a distance-invariant measure of path shape similarity, and a distance-sensitive metric invariant to shape.

In the case of tracking data from 2-d image, there are two dimensions to the data which need comparing. The tests can be run separately in each dimension, and the results combined. This can also provide extra information, as error in one dimension only might suggest an interesting phenomenon in the tracking that only occurs along one axis of the image.

4. Failure Modes

The statistical measures proposed above provide well-defined comparative tests which can be applied to the output of competing algorithms or to ground truth and tracker output. The analysis provided is at the level of the test data set, in the case of McNemar’s statistic, and the individual tracker path, in the case of the proposed permutation tests. Though clearly valuable, this kind of analysis will invariably be supported by closer inspection of tracker behaviour. To identify the failure modes of a proposed algorithm it is necessary to look in detail at selected sections of the tracker’s path, usually around points at which tracking is lost. The aim of this examination is to determine exactly why the algorithm failed, an activity closely related to debugging. For example, it might become clear that tracking always fails when a target passes over a certain area of the image, or when the target makes a certain kind of movement.

Detailed examination of failure modes effectively involves segmenting the tracker’s path into areas of success and failure, by comparing it with the available ground truth. This is usually done by eye, and so is both time-consuming and prone to subjectivity.

4.1. The Tool

To meet this need, an automated tool has been created based upon the permutation tests described above. The aim is to support fast and objective analysis of the success or failure of a tracker when comparing its output to ground truth. This analysis can be achieved by segmenting the tracker output, splitting its path into progressively smaller sections, and running appropriate accuracy tests (above) on these subsets.

Segmentation can be performed in a variety of ways. The algorithm employed might take into account the distance between tracker and ground truth data at each time point, employing some form of thresholding to identify breakpoints. Alternatively, it might segment recursively, applying statistical tests first to the entire track then, as the track is segmented, to a set of increasingly small sub-sections. Sections of track
which passed the applied tests would be left unsegmented, while those which failed the tests would be subdivided until some stopping criterion was met. For the present, to provide a demonstration of concept and allow initial experimentation, a simple binary chop approach has been used to segment the tracks (see Figure 2).

**Figure 2. Tracker evaluation using a binary chop segmentation algorithm**

```
BINARY CHOP TRACKER EVALUATION
on dataset of size n
if spearman permutation test pass and
mean permutation test pass or
n/2 <= maximum depth
return current node
else
    current node left child = run algorithm
    on first half of n
    current node right child = run algorithm
    on second half of n
return current node including descendants
end if
```

4.2. Results

The remainder of this section presents results obtained when applying the binary chop tool incorporating the proposed path shape and distance tests to two image sequences used in the evaluation of the KAMS tracking algorithm [6]. For each sequence, the KAMS [6] and Annealed particle filter [7] algorithms are compared to a manually marked ground truth. It should be stressed, however, that the examples chosen here are intended merely to illustrate the operation of the proposed tool, and not to provide any systematic evaluation of the tracking algorithms used.

Figure 3a shows the first frame of an image sequence depicting a sparsely populated rail station. Overlaid on this image is the path taken by the annealed particle filter [6] when applied to one individual. The path is colour-coded; green where all tests (permutation tests using mean and Spearman tests on both the x and y coordinate plots) are passed, amber where one test is passed and the other failed, and red where both spearman and mean tests fail on at least one coordinate.

Displays of the type shown in Figure 3a give an immediate impression of the given tracker’s performance. Figures 3b and c show colour coded plots of the tracker’s and ground truth estimates of x (Figure 3b) and y (Figure 3c) coordinates. Here the colour of the data plot distinguishes tracker (light) from ground truth (darker). The results of the statistical tests are shown in the background colour, coded as before. This type of display allows more detailed examination of the results.

Note first of all that there is no green on the path; at each point some test has failed. Figure 3b shows the tracker to have failed completely some 35% of the way through the sequence, and marks this event clearly. The y-coordinate data in Figure 3c is meaningless beyond this point, but also shows errors in the early part of the sequence, where the x coordinate also fails the mean test. The tracker had begun to drift off the target before it failed completely.

Closer examination also shows that, while remaining associated with the correct target throughout, the tracker does tend to drift up and down the tracked individual’s body, as might be expected from an algorithm employing the type of colour histogram representation used here (see [6] for implementation details).
Figure 3. (a) 2D display of the output of the annealed particle filter. The colour of the line indicates which combination of statistical tests were passed during each segment. (b) annealed PF output versus groundtruth, X. (c) annealed PF output versus groundtruth, Y. Time is increasing along the x-axis, and the y-axis corresponds to the coordinate in the image plane. See text for colour coding.

Figure 4. (a) 2D display of the KAMS tracking output. The colour of the line indicates which combination of statistical tests were passed during each segment. (b) KAMS output versus groundtruth, X. (c) KAMS output versus groundtruth, Y. Time is increasing along the x-axis, and the y-axis corresponds to the coordinate in the image plane. See text for colour coding.
Figure 4 shows the output of the KAMS tracker given the same sequence. The x-values satisfy all the tests throughout the sequence, but the y-values are less reliable, and fail both mean and Spearman tests on two occasions. Note that the plot is Figure 4c is quite well segmented into regions in which the tracker output and ground truth separate, but follow similarly-shaped paths and areas in which the shape of the path also changes.

Figure 5 shows data obtained by application of the annealed particle filter [7] to an image sequence of an underground railway platform. Figure 5b neatly illustrates the analysis performed by the evaluation tool. Moving from left to right we see a period during which the x-coordinate passed both tests, followed by one in which the slightly harsh mean test failed. As the tracker and ground truth diverge, both tests fail; the two paths are both a different shape and offset from one another. In the central region the Spearman test is passed, as the two paths are of similar shape, but the mean test fails as the tracker moves alongside, not on top of, the target. As the two paths come together again, both tests fail, as they did earlier.

Perhaps the most striking feature of this data, however, is that large parts of the path fail both tests in the y-coordinate. This is due in part to the colour histogram used to represent the target and the elongated nature of the (human) targets, but also reflects the hard and fixed criteria set by the statistical tests. The mean test in particular can be harsh, particularly if the distributions compared are tightly clustered around their means. This can be softened if desired by changing the significance level, but, as x-coordinate data typically passed the default mean test in the experiments conducted to date, for the present we have resisted this temptation.

As with any results, further manual inspection is still necessary. For example, the 2d output shown in Figure 5a indicates that some of the errors might be cause by the kerb near the railway tracks, or the cluttered background near the pillar. However, it is only on watching the video that it becomes clear that it is in fact another person causing the tracking error, not the background. However, outputs such as figure 5 still give a useful indication of where in the frame the event occurs, guiding the manual observer and making it easy to identify significant failure modes.

![Figure 5a](image1.png)

![Figure 5b](image2.png)

![Figure 5c](image3.png)

**Figure 5.** (a) 2D display of the output of an annealed particle filter. The colour of the line indicates which combination of statistical tests were passed during each segment. (b) tracker output versus groundtruth, X. (c) tracker output versus groundtruth, Y. Time is increasing along the x-axis, and the y-axis corresponds to the coordinate in the image plane. See text for colour coding.
6. Conclusion

This paper has argued for the use of a set of statistical tests in the assessment of the robustness and accuracy of visual tracking algorithms. These tests are used to give quantitative results where qualitative methods are commonly used, and have been built into a tool to highlight to the user which sections of the tracked paths contain which kind of errors. This information can be presented in a meaningful way in the software tool, both overlaid on the 2D image plane and displayed overlaid onto traditional-style coordinate graphs. In particular:

- McNemar’s statistic has been applied to tracking results to give a statistically valid comparison of the robustness of different algorithms.

- Novel ways of measuring the accuracy of tracking compared to a groundtruth have been proposed, using statistical measures of path shape and distribution distance. These measures can be applied to sub-sections of tracked paths to determine the kind of errors present at different points in the tracking.

- A tool has been built to generate intuitive visual output, to aid further manual analysis of the video results. This tool can display the results of the statistical tests both spatially over an image frame and over time in coordinate graphs. Applying the tests to the x- and y-axis of the image separately can add further insight as to the cause of the error. These visual results can be quickly interpreted and used to select points of interest for further manual investigation from the video output itself.

Future work will focus on examining the role and value of the proposed measurements and prototype tool in real performance evaluations. In particular, consideration will be given to the potential use of this kind of method in multi-target tracking applications; at present each target/tracker pair would have to be considered independently. Technical developments will centre on the use of alternative, and possibly more accurate segmentation techniques within the evaluation tool.

6. References


