University of Leeds
SCHOOL OF COMPUTING
RESEARCH REPORT SERIES
Report 2003.01

Performance Evaluation Metrics and Statistics for Positional Tracker Evaluation

by

C J Needham & R D Boyle

January 2003

Abstract

This report discusses methods behind tracker evaluation, the aim being to evaluate how well a tracker is able to determine the position of a target object. Few metrics exist for positional tracker evaluation; here the fundamental issues of trajectory comparison are addressed, and metrics are presented which allow the key features to be described. Often little evaluation on how precisely a target is tracked is presented in the literature, with results detailing for what percentage of the time the target was tracked. This issue is now emerging as a key aspect of tracker performance evaluation.

The metrics developed are applied to real trajectories for positional tracker evaluation. Data obtained from a sports player tracker on video of a 5-a-side soccer game, and from a vehicle tracker, is analysed. These give quantitative positional evaluation of the performance of computer vision tracking systems, and provides a framework for comparison of different methods and systems on benchmark data sets.

1 Introduction

There are many ways in which the performance of a computer vision system can be evaluated. Often little evaluation on how *precisely* a target is tracked is presented in the literature, with the authors tending to say for what percentage of the time the target was tracked. This problem is beginning to be taken more seriously, and an annual workshop on performance evaluation of tracking and surveillance [5] has begun recently (2000).

Performance evaluation is a wide topic, and covers many aspects of computer vision. Ellis [1] discusses approaches to performance evaluation, and covers the different areas, which include how algorithms cope in different physical conditions in the scene, i.e. weather, illumination and irrelevant motion, to assessing performance through ground truthing and the need to compare tracked data to marked up data, whether this be targets' positions, 2D shape models, or classification of some description.

In previous work [4] , mean and standard deviations of errors in tracked data from manually marked up data has been presented, with simple plots. Harville [2] presents similar positional analysis when evaluating the results of person tracking using plan-view algorithms on footage from stereo cameras. In certain situations Dynamic Programming can be applied to align patterns in feature vectors, for example in the speech recognition domain as Dynamic Time Warping (DTW) [6]. In this work trajectory evaluation builds upon comparing equal length trajectories having frame by frame time steps with direct correspondences.

When undertaking performance evaluation of a computer vision system, it is important to consider the requirements of the system. Common applications include detection (simply identifying if the target object is present), coarse tracking (for surveillance applications), tracking (where reasonably accurate locations of target objects are identified), and high-precision tracking (for medical applications, reconstructing 3D body movements). This report focuses on methods behind *positional* tracker evaluation, the aim being to evaluate how well a tracker is able to determine the position of a target object, for use in tracking and high-precision tracking as described above.

2 Metrics and statistics for trajectory comparison

Few metrics exist for positional tracker evaluation. In this section the fundamental issues of trajectory comparison are addressed, and metrics are presented which allow the key features to be described. In the following section, these metrics are applied to real trajectories for positional tracker evaluation.

2.1 Trajectory definition

A **trajectory** is a sequence of positions over time. The general definition of a trajectory T is a sequence of positions (x_i, y_i) and corresponding times, t_i:

$$T = \{(x_1, y_1, t_1), (x_2, x_2, t_2), \ldots, (x_n, y_n, t_n)\} \tag{1}$$

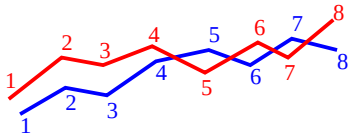


Figure 1: Example of a pair of trajectories.

In the computer vision domain, when using video footage, time steps are usually equal, and measured in frames. Thus, t_n may be dropped, as the subscript on the positions can be taken as time, and Equation 1 becomes:

$$T = \{(x_1, y_1), (x_2, x_2), \ldots, (x_n, y_n)\} \tag{2}$$

i.e. trajectory T is a sequence of (x_i, y_i) positions at time step i, as illustrated in Figure 1. **Paths** are distinguished from trajectories by defining a path as a trajectory not parameterised by time.

To evaluate the performance of the tracker, metrics comparing two trajectories need to be devised. We have two trajectories T_A and T_B which represent the trajectory of a target from the tracker, and the ground truth trajectory - which is usually marked up manually from the footage. Metrics comparing the trajectories allow us to identify how *similar*, or how *different* they are.

2.2 Comparison of trajectories

Consider two trajectories composed of 2D positions at a sequence of time steps. Let positions on trajectory T_A be (x_i, y_i), and on trajectory T_B be (p_i, q_i), for each time step i. The displacement between positions at time step i is given by $\mathbf{d}_i$:

$$\mathbf{d}_i = (p_i, q_i) - (x_i, y_i) = (p_i - x_i, q_i - y_i) \tag{3}$$

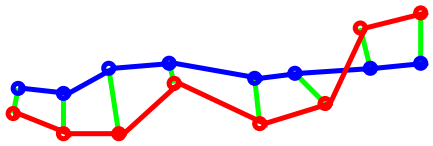


Figure 2: Comparison of displacement between two trajectories.

And the distances between the positions at time step i are given by d_i:

$$d_i = |\mathbf{d}_i| = \sqrt{(p_i - x_i)^2 + (q_i - y_i)^2} \tag{4}$$

A metric commonly used for tracker evaluation is the mean of these distances [4, 2]. We shall call this metric m_1.

$$m_1 = \mu(d_i) = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{5}$$

m_1 gives the average distance between positions at each time step. Figure 2 shows two trajectories and identifies the distance between corresponding positions. The distribution of these distances is also of significance, as it shows how the distances between trajectories (tracker error) are spread, as illustrated in Figure 3, where a skewed distribution can be seen.

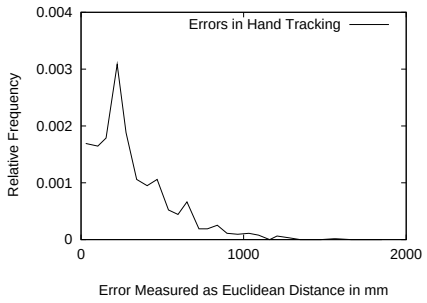


Figure 3: Distribution of distances between positions.

Other statistics provide quantitative information about the distribution. Here we identify the mean, median (expected to be lower than the mean, due to the contribution to the mean of the furthest outliers), standard deviation, minimum and maximum values as useful statistics for describing the data. Let us define $\mathcal{D}(T_A, T_B)$ to be the set of distances d_i between trajectory A and B. The above statistics can be applied to this set:

$$
\begin{array}{lrcll}
\text{Mean} & \mu(\ \mathcal{D}(T_A, T_B)\ ) & = & \frac{1}{n}\sum_{i=1}^{n} d_i & \\
\text{Median} & median(\ \mathcal{D}(T_A, T_B)\ ) & = & d_{\frac{n+1}{2}} & \text{if } n \text{ odd,} \\
 & & = & \frac{1}{2}(d_{\frac{n}{2}} + d_{\frac{n}{2}+1}) & \text{if } n \text{ even} \\
\text{Standard deviation} & \sigma(\ \mathcal{D}(T_A, T_B)\ ) & = & \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \mu(d_i))^2} & \\
\text{Minimum} & min(\ \mathcal{D}(T_A, T_B)\ ) & = & \text{the smallest } d_i & \\
\text{Maximum} & max(\ \mathcal{D}(T_A, T_B)\ ) & = & \text{the largest } d_i &
\end{array}
\tag{6}
$$

2.3 Spatially separated trajectories

Some pairs of trajectories may be very similar, except for a constant difference in some spatial direction (Figure 4). Defining a metric which takes this into account may reveal a closer relationship between two trajectories.

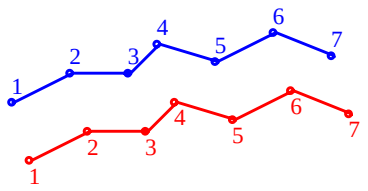


Figure 4: Two spatially separated trajectories.

Given the two trajectories T_A and T_B, it is possible to calculate the optimal spatial translation $\hat{\mathbf{d}}$ (shift) of T_A towards T_B, for which m_1 is minimised. $\hat{\mathbf{d}}$ is the average displacement between the trajectories, and is calculated as:

$$\hat{\mathbf{d}} = \mu(\mathbf{d}_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{d}_i \tag{7}$$

Now we can define $\mathcal{D}(T_A + \hat{\mathbf{d}}, T_B)$ to be the set of distances between a translated trajectory T_A (by $\hat{\mathbf{d}}$) and T_B. The same statistics can be applied to this set, $\mathcal{D}(T_A + \hat{\mathbf{d}}, T_B)$, to describe the distances. $\mu(\ \mathcal{D}(T_A + \hat{\mathbf{d}}, T_B)\ ) < \mu(\ \mathcal{D}(T_A, T_B)\ )$ in all cases, except when the trajectories are already optimally spatially aligned.

When $\mu(\ \mathcal{D}(T_A + \hat{\mathbf{d}}, T_B)\ )$ is significantly lower than $\mu(\ \mathcal{D}(T_A, T_B)\ )$, it may highlight a tracking error of a consistent spatial difference between the true position of the target, and the tracked position.

2.4 Temporally separated trajectories

Some pairs of trajectories may be very similar, except for a constant time difference (Figure 5). Defining a metric which takes this into account may reveal a closer relationship between two trajectories.

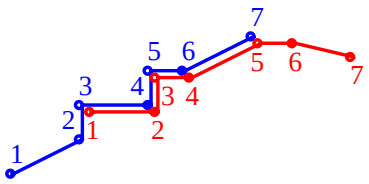


Figure 5: Two temporally separated trajectories.

Given the two trajectories T_A and T_B, it is possible to calculate the optimal temporal translation j (shift) of T_A towards T_B, for which m_1 is minimised. When the time-shift j is positive $T_{A,i}$ is best paired with $T_{B,i+j}$, and when j is positive $T_{A,i+j}$ is best paired with $T_{B,i}$. Time-shift j is calculated as:

$$j = \arg\min_k \Big(\frac{1}{n-|k|} \sum_{i=Q}^{R} |(p_{i+k}, q_{i+k}) - (x_i, y_i)| \Big)$$
$$\text{if } k \geqslant 0 \text{ then } Q = 0 \text{ else } Q = -k. \;\; R = Q + n - |k|. \tag{8}$$

Now we can define $\mathcal{D}(T_A, T_B, j)$ to be the set of distances between a temporally translated trajectory T_A or T_B, depending on j's sign. The same statistics as before can be applied to this set, $\mathcal{D}(T_A, T_B, j)$, to describe the distances. $\mu(\ \mathcal{D}(T_A, T_B, j)\ ) < \mu(\ \mathcal{D}(T_A, T_B)\ )$ in all cases, except when the trajectories are already optimally temporally aligned.

When $\mu(\ \mathcal{D}(T_A, T_B, j)\ )$ is significantly lower than $\mu(\ \mathcal{D}(T_A, T_B)\ )$, it may highlight a tracking error of a consistent temporal difference between the true position of the target, and the tracked position. In practice j should be small; it may highlight a lag in the tracked position (Figure 5).

2.5 Spatio-Temporally separated trajectories

Combining the spatial and temporal alignment process identifies a fourth distance statistic. We define $\mathcal{D}(T_A + \hat{\mathbf{d}}', T_B, j)$ to be the set of distances between the spatially and temporally optimally aligned trajectories, where $\hat{\mathbf{d}}' = \hat{\mathbf{d}}(T_A, T_B, j)$ is the optimal spatial shift between the temporally shifted (by j time steps) trajectories.

The procedure for defining this set is similar to above; calculate the optimal j for which the mean distance between space (translation of $\hat{\mathbf{d}}'$) and time (time-shift of j) shifted positions is minimised, using an exhaustive search. Once j has been calculated, the set of distances $\mathcal{D}(T_A + \hat{\mathbf{d}}', T_B, j)$ can be formed, and the usual statistics can be calculated.

When the trajectories are spatio-temporally aligned, the mean value, $\mu(\ \mathcal{D}(T_A + \hat{\mathbf{d}}', T_B, j)\ )$ is less than or equal to the mean value of the three other sets of distances; when the trajectories are unaltered, spatially aligned, or temporally aligned.

2.6 Area between trajectories

The area between two trajectories provides time independent information. The trajectories must be treated as paths whose direction of travel is known.

Given two paths A and B, the area between them is calculated by firstly calculating the set of crossing points where path A and path B intersect. These crossing points are then used to define a set of regions. If a path crosses itself *within* a region, then the loop created is discarded by deleting the edge points on the path between where the path crosses itself. This resolves the problem of calculating the area if a situation where a path crosses itself many times occurs, as illustrated in Figure 6. Now the area between

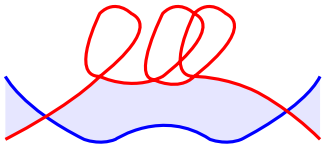


Figure 6: Regions with self crossing trajectories. The shaded regions show the area calculated.

the paths can be calculated as the summation of the areas of the separate regions. The area of each region is calculated by treating each region as an n-sided polygon defined by edge points (x_i, y_i) for $i = 1, \ldots, n$, where the first point is the intersection point, the next points follow those on path A, then the second crossover point, back along path B to the first point. i.e. the edge of the polygon is traced. Tracing the polygon, the area under each edge segment is calculated as a trapezoid, each of these is either added to or subtracted from the total, depending on its sign, which results from the calculation of $(x_{i+1} - x_i)(y_i + y_{i+1})/2$ as the area between the x-axis and the edge segment from (x_i, y_i) to (x_{i+1}, y_{i+1}). After re-arrangement Equation 9 shows the area of such a region. (It does not matter which way the polygon is traced, since in our computation the modulus of the result is taken).

$$A_{region} = \left| \frac{1}{2} \left(\left((\sum_{i=1}^{n-1} x_{i+1} y_i) + x_1 y_n \right) - \left((\sum_{i=1}^{n-1} x_1 y_{i+1}) + x_n y_1 \right) \right) \right| \qquad (9)$$

The areas of each of the regions added together gives the total area between the paths, and has dimensionality L^2 i.e. mm^2. To obtain a useful value for the area metric, the area calculated is normalised by the average length of the paths. This gives the 'area' metric on the same scale as the other distance statistics. It represents the average time independent distance (in mm) between the two trajectories, and is a **continuous average distance**, rather than the earlier discrete average distance.

3 Evaluation, results and discussion

Performance evaluation is performed on two tracking systems; a sports player tracker [4] , and a vehicle tracker [3]. Figure 7 shows example footage used in each system. First, the variability between two hand marked up trajectories is discussed.

Figure 7: Example footage used for tracking.

3.1 Comparison of two hand marked up trajectories

This section compares two independently hand marked up trajectories of the same soccer player during an attacking run (both marked up by the same person, by clicking on the screen using the mouse). There are small differences in the trajectories, and they cross each other many times. The results are shown in Table 1, and the trajectories are

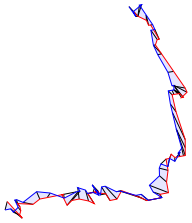


Figure 8: Two example hand marked up trajectories, showing the area between them, and the displacements between positions at each time step.

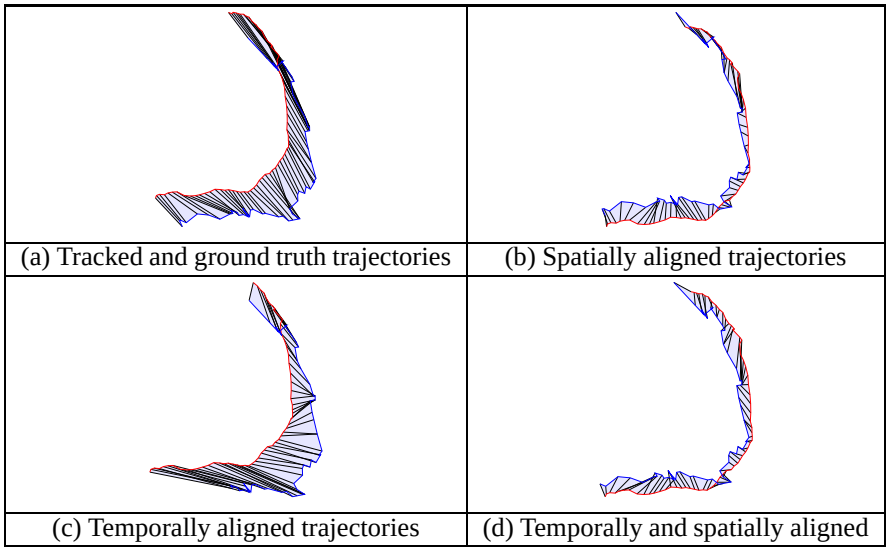
shown graphically in Figure 8 with the area between the two paths shaded, and dark lines connecting the positions on the trajectories at each time step. The second row of Table 1 identifies an improvement in the similarity of the two trajectories if a small spatial shift of $\hat{\mathbf{d}} = (-58, 74)$ in mm, is applied to the first trajectory. As expected in hand marked up data, the two trajectories are optimally aligned in time (time-shift $j = 0$).

Metric	mean	median	min	max	s.d	'area'
$\mathcal{D}(T_A, T_B)$	134	115	0	444	89	56
$\mathcal{D}(T_A + (-55, 74), T_B)$	110	92	10	355	72	42
$\mathcal{D}(T_A, T_B, 0)$	134	115	0	444	89	56
$\mathcal{D}(T_A + (-55, 74), T_B, 0)$	110	92	11	355	72	42

Table 1: Results of trajectory evaluation. All distances are in mm.

3.2 Sports player tracker example

This section compares a tracked trajectory, T_C, to a hand marked up trajectory T_B. The sports player tracker [4] identifies the ground plane position of the players, which is taken as the mid-point of the base of the bounding box around the player, and is generally where the players' feet make contact with the floor. Figure 9 qualitatively illustrates the shifted trajectories, whilst Table 2 quantitatively highlights the systematic error present in this sequence.



(a) Tracked and ground truth trajectories | (b) Spatially aligned trajectories

(c) Temporally aligned trajectories | (d) Temporally and spatially aligned

Figure 9: (a)-(d) Example trajectories over 70 frames. Trajectory T_C from tracker compared to T_B - the hand marked up trajectory. The figures show the area between them, and the displacements between positions at each time step.

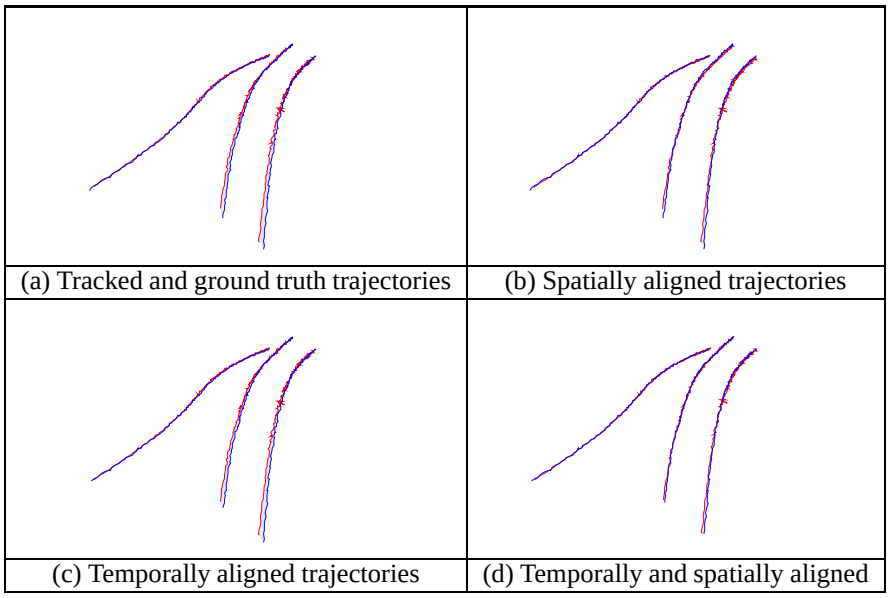
If T_C is shifted by 500mm in the x-direction, and $600 - 700$mm in the y-direction, the differences between the trajectories fall significantly. This may be due to an invalid assumption that the position of the tracked players is the mid-point of the base of the bounding box around the player. This may be due to the player's shape in these frames, tracker error, or human mark up of the single point representing the player at each time step.

Metric	mean	median	min	max	s.d	'area'
$\mathcal{D}(T_C, T_B)$	890	859	393	1607	267	326
$\mathcal{D}(T_C + (510, -710), T_B)$	279	256	40	67	145	133
$\mathcal{D}(T_C, T_B, -9)$	803	785	311	1428	237	317
$\mathcal{D}(T_C + (551, -618), T_B, -2)$	263	230	52	673	129	138

Table 2: Results of trajectory evaluation. All distances are in mm.

3.3 Car tracker example

This section compares trajectories from a car tracker [3] with manually marked up ground truth positions. In this example, the evaluation is performed in image plane coordinates (using 352×288 resolution images), on a sequence of cars on an inner city bypass, a sample view is shown in Figure 7.



(a) Tracked and ground truth trajectories | (b) Spatially aligned trajectories

(c) Temporally aligned trajectories | (d) Temporally and spatially aligned

Figure 10: (a)-(d) Three pairs of example trajectories over 200 frames. Trajectory T_A from tracker compared to T_B - the hand marked up trajectory, with the area between them shaded.

Trajectory comparison is performed on three trajectories of cars in the scene, each over 200 frames in length. Figure 10 displays these trajectories along with the ground truth, and Table 3 details the quantitative results, from which it can be seen that there is little systematic error in the system, with each car's centroid generally being accurate to between 1 and 3 pixels.

Left Path

Metric	mean	median	min	max	s.d	'area'
$\mathcal{D}(T_A, T_B)$	1.7	1.3	0.1	7.1	1.3	0.6
$\mathcal{D}(T_A + (-0.2, -0.8), T_B)$	1.6	1.3	0.2	6.5	1.1	0.5
$\mathcal{D}(T_A, T_B, 1)$	1.5	1.3	0.1	5.1	0.8	0.6
$\mathcal{D}(T_A + (0.8, -0.1), T_B, 1)$	1.3	1.2	0.1	5.2	0.8	0.5

Middle Path

Metric	mean	median	min	max	s.d	'area'
$\mathcal{D}(T_A, T_B)$	3.0	2.3	0.4	12.4	2.2	1.8
$\mathcal{D}(T_A + (1.9, -0.9), T_B)$	2.3	1.9	0.1	11.2	2.0	0.9
$\mathcal{D}(T_A, T_B, 1)$	2.9	2.3	0.5	8.7	1.4	1.8
$\mathcal{D}(T_A + (3.1, 1.8), T_B, 3)$	1.3	1.3	0.1	3.6	0.7	0.6

Right Path

Metric	mean	median	min	max	s.d	'area'
$\mathcal{D}(T_A, T_B)$	3.2	2.9	0.3	9.7	1.8	2.1
$\mathcal{D}(T_A + (2.3, -0.2), T_B)$	2.5	2.3	0.1	8.6	1.4	1.2
$\mathcal{D}(T_A, T_B, 0)$	3.2	2.3	0.3	9.7	1.8	2.1
$\mathcal{D}(T_A + (2.9, 2.0), T_B, 2)$	1.7	1.6	0.1	6.0	0.9	1.0

Table 3: Results of trajectory evaluation. All distances are in pixel units.

4 Summary and conclusions

Quantitative evaluation of the performance of computer vision systems allows their comparison on benchmark datasets. It must be appreciated that algorithms can be evaluated in many ways, and we must not lose target of the aim of the evaluation. Here, a set of metrics for positional evaluation and comparison of trajectories has been presented. The specific aim has been to compare two trajectories. This is useful when evaluating the performance of a tracker, for quantifying the effects of algorithmic improvements. The spatio/temporally separated metrics give a useful measure for identifying the precision of a trajectory, once the systematic error is removed, which may be present due to a time lag, or constant spatial shift. There are many potential obvious uses for trajectory comparison in tracker evaluation, for example comparison of a tracker with Kalman Filtering and without [4] (clearly this affects any assumption of independence).

It is also important to consider how accurate we require a computer vision system to be (this may vary between detection of a target in the scene and precise location of a targets' features). Human mark up of ground truth data is also subjective, and there are differences between ground truth sets marked up by different individuals. If we require a system that is at least as good as a human, in this case, the tracked trajectories should be compared to how well humans can mark up the trajectories, and a statistical test performed to identify if they are significantly different.

References

[1] T. J. Ellis. Performance metrics and methods for tracking in surveillance. In *3rd IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, Copenhagen, Denmark, 2002.

[2] M. Harville. Stereo person tracking with adaptive plan-view statistical templates. In *Proc. ECCV Workshop on Statistical Methods in Video Processing*, pages 67–72, Copenhagen, Denmark, 2002.

[3] D. R. Magee. Tracking multiple vehicles using foreground, background and motion models. In *Proc. ECCV Workshop on Statistical Methods in Video Processing*, pages 7–12, Copenhagen, Denmark, 2002.

[4] C. J. Needham and R. D. Boyle. Tracking multiple sports players through occlusion, congestion and scale. In *Proc. British Machine Vision Conference*, pages 93–102, Manchester, UK, 2001.

[5] IEEE Workshop on Performance Evaluation of Tracking and Surveillance (2000). `http://visualsurveillance.org/PETS2000` Last accessed: 18/10/02.

[6] H. Sakoe and S. Chiba. Dynamic Programming optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.