Performance Evaluating the Evaluator

Thor List, José Bins, Jose Vazquez and Robert B. Fisher
School of Informatics, University of Edinburgh, UK
thor.list@ed.ac.uk

Abstract

When evaluating the performance of a computer-based visual tracking system one often wishes to compare results with a standard human observer. It is a natural assumption that humans fully understand the relatively simple scenes we subject our computers to and because of this, two human observers would draw the same conclusions about object positions, tracks, size and even simple behaviour patterns. But is that actually the case?
This paper will provide a baseline for how computer-based tracking results can be compared to a standard human observer.

1. Introduction

Automated visual surveillance and tracking systems are being developed and deployed in several countries, and many universities have research groups working on improving them, in terms of performance and efficiency. The performance is usually compared to a standard human observer, in the form of evaluation training sets which have been pre-analysed and annotated by hand by a human observer. Whole conferences and workshops (e.g. the PETS series) are held where participants show their results of this comparison, and often results are compared between research groups by assuming that the human observer is equal and infallible. But this is clearly not the case and human variability needs to be taken into account when comparing results to the ground truth. For example, if the sizes of the bounding boxes provided by two human observers vary by 5%, then a difference of 5% between a tracker result and the 'ground truth' is probably meaningless.

With the emergence of standard tools for creating such annotation, or 'ground truth', it has become easier to produce and share these annotations. However, having standard tools also provides the means to test the so-called standard human observer assumption, by having more than one person use the same tool to annotate the same video sequences.

This paper evaluates consistency in the ground truth video sequence labelling produced by the CAVIAR project, to give an example of expected level of variation in video sequence annotation.

The CAVIAR project has produced 90 hand-labelled video sequences (for a total of about 90K frames) of people interacting in an office lobby and a shopping centre. The labelling covers bounding boxes of the moving individuals and also a semantic description of their behaviour.

A summary of the first third of these data sets is found at [4]. The ground truth labelling was produced by a interactive JAVA-based tool for drawing bounding boxes around people and buttons for setting symbolic descriptions. This tool is similar to the ViPER tool [6,3].

One sequence of 958 frames was labelled by three different people, and reviewed for correctness by a fourth person. We use this sequence as the basis for the evaluation presented in this paper.

The three sets of ground truth were produced by two PhD and one graduating undergraduate students. They were instructed to mark all moving people, according to written and oral instructions. The ground truth was checked automatically for semantic correctness and all three labellings were reviewed by one supervisor to improve consistency. All three people had some computer vision background. The sequence labelled was the "Fight_One_ManDown" sequence. See this setting at http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/. The main activities observed are several people standing around idly or walking, plus a simulated fight between two researchers. The sequence also includes some difficult cases, such as people barely moving, people in ambiguous situations and small targets in dark regions at the rear of the scene. In doing the labelling, observers knew about various scene furniture, such as information points, desks, couches, etc.

1.1 Data Model

Each individual person is described by a bounding box (id, centre coordinates, width, height, orientation of main axis of individual). Individuals are only labelled once they start moving; otherwise they are effectively background.

The symbolic labelling of the people is based on a behaviour model proposed by Crowley et al [2]. In this model, a *situation* is a short term activity, such as stopping to look in a shop window. Each individual has a *role* in the current situation, i.e. a window-shopper.

The overall *context* describes a network of situations that the individuals may pass through in the course of a longer term activity, for example a window-shopping expedition.

Based on this semantics of the activity interpretation, each individual is labelled with a role (e.g. fighter, browser, left victim, leaving group, walker, left object), is a participant in a situation (e.g. browsing, moving, inactive), which is a component of a context (e.g. Walking, Idleness, Browse, Collapse, Leaving object, Meeting, Fighting).

Each individual is labelled in each frame with one member of each of the above sets of labels. The semantics of activity labelling is constrained by a finite-state model of the allowable behaviours, which define the allowable sequences of situations in a given context. An example graph for a loitering context is:

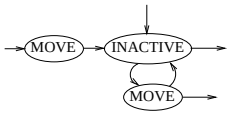

Figure 1. Sequence graph for "idleness" (aka loitering)

As well as the role, the ground truth labelling for the box has a qualitative assessment of the activity level of the individual or group, i.e. whether they are running, walking, stationary but active (e.g. moving arms), or inactive.

Each video frame contains zero or more labelled individual or group boxes. The boxes are labelled with an identifier, which persists as long as the individual is visible. If a person disappears and then later reappears, then the individual obtains a new identity. If the person is obscured/occluded for only a few frames, then the same identity is maintained.

The ground truth is encoded in CVML, an XML specialization for computer vision applications [5].

1.2 Statistics Computed

In this paper we report on the statistical variation in the hand labelling. This variation is at 3 levels:

1. Geometric description of the moving individuals: the bounding box positions and sizes and the main elongation axis of the person.
2. Differences in the choice of activity, role, situation and context labels.
3. Differences in the frame times at which a given activity, role, situation and context starts and ends.

1.3 Previous work on Evaluating Consistency in Ground Truth Labelling

We are not aware of any consistency studies concentrating explicitly on video sequence analysis, but know of studies on tissue boundary tracing in MRI [1] and single image region segmentation [7].

Crawford-Hines [1] presents a method for learning boundary models from demonstration labellings by experts, rather than by a priori assumed image boundary properties. As part of this study, the same expert twice traced the boundary around the same structure. The main statistic measured was the average distance between the two curves, along with a cumulative distribution of distances (e.g. less than 1 pixel 86% of the time).

Martin [7] looked at the semantic region segmentation of everyday images (e.g. people, animals, cars, buildings, etc) as part of a study on grouping and segmentation. In this case, the main effect was the different choices in details to be segmented by the humans. He concluded that a good model for explaining the observed differences was to assume that there was an underlying tree of segmentations, perhaps down to the individual pixel, and different people pruned the branches of that tree at different levels. On the basis of his refinement-based model, he defined measures that tested to what extent two segmentations were consistent, as a function of region overlap.

Martin et al [8] also exploited human variability in a learning-based colour, brightness and texture region segmentation algorithm. Their algorithm learns from a large set of human segmentations and the various segmenters are compared to a nominal top performance derived from the range of human segmentations.

Some of the issues raised are also relevant to sequence semantic labelling (independence of sampling rates, robustness to slight variations at segmentation boundaries, underlying implicit model of behaviour).

Thus, although there are not actual studies on sequence labelling, the methods of cumulative difference distributions and containment within an underlying implicit model are relevant here.

2. Spatial Tracking

The geometric descriptions of tracked individuals and groups are annotated by the human observers using a 2D axis aligned bounding box that fully surrounds the object or group of objects. When comparing the output from a tracker with the ground truth measurements such as true, false and missed detections are computed, as well as accuracy in position and size.

We compared the performance of our three human observers by vetting Observers 2 and 3 against Observer 1. In this section we found symmetrical results when reversing the comparisons (1 and 3 against 2, etc.).

2.1 Object Detection

We first measure how well Observers 2 and 3 agreed with Observer 1 overall on object detection. In Figure 2 we show true detections versus overlap requirement of Observers 2 and 3 as compared with Observer 1. The overlap percentage is the ratio of bounding box intersection to the box of Observer 1. We see that when only requiring the bounding boxes to overlap 60% or less Observer 2 detects ~98% of all boxes, whereas Observer 3 detects ~95%. We conclude that detection is accurate at about the 95% level.

Observer 3 overall missed a few boxes as there are some people in the scene whose activity level is subjective and hence may be treated as stationary or non-detectable. When requiring a 90% or better precision in detection, we see that Observers 2 and 3 both agree very poorly with Observer 1. This is explained by the lack of accuracy of the bounding boxes; both in position (see Figure 6) and in size (see Figure 8). These two effects combined often push the overlap below 90%.

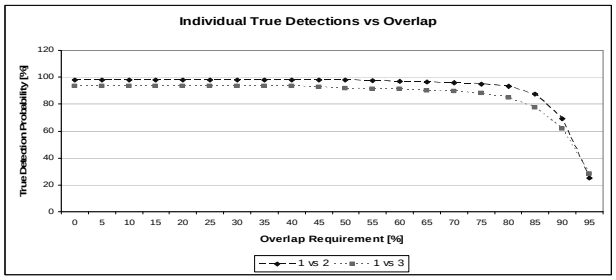

Figure 2. Individual true detections versus overlap

False detections happen when Observers 2 or 3 find objects that Observer 1 did not mark with a box. This can also happen when the objects' bounding boxes do not match more than the required overlap. The individual false detections are shown in Figure 3. Here we see that Observer 3 generally marked 20% more objects than Observer 1 and Observer 2 a surprising 80% more objects than Observer 1. This is because Observer 2 detected a nearly stationary person. Observers 1 and 3 did not, as since these people are in the scene the majority of the time, they add up to the 80% false detection rate. The reason for these vastly different results lies in the human observers different opinions regarding object saliency.

2.2 Group detection

The observers were also asked to mark groups of individuals when it was clear that two or more individuals were interacting. Although one would assume that this definition would cause more confusion than the individual detections above, one can see from Figure 4 that all three observers agree quite well on what constitutes a group, up until 90% or more overlap is required.

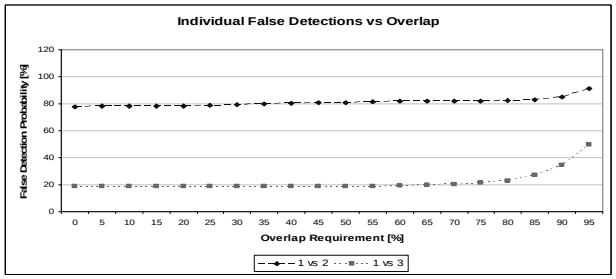

Figure 3. Individual false detections versus overlap

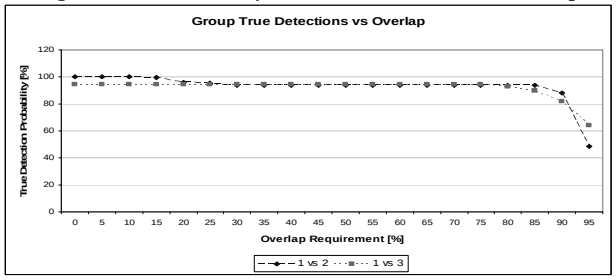

Figure 4. Group true detections versus overlap

As seen in Figure 5 Observer 3 agrees very well with Observer 1 as there are practically no false group detections, whereas Observer 2 groups two independent people that are walking towards each other.

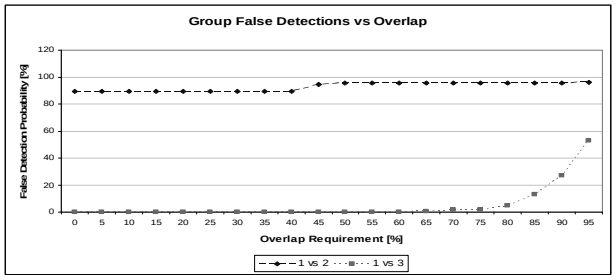

Figure 5. Group false detections versus overlap

2.3 Position and size

When comparing the positions and sizes of an object's bounding box one is evaluating the precision with which the observers have marked the location and apparent size of an agreed detected object. Figures 6 to 13 should ideally show error bars, but adding these would clutter the graphs. We observed that the error is roughly constant. For each figure we have measured how much 1 standard deviation is at 50% from the total number of data points, N = 3568 (1 vs 2) and N = 3348 (1 vs 3) for individuals, and N = 166 (1 vs 2) and N = 167 (1 vs 3) for groups.

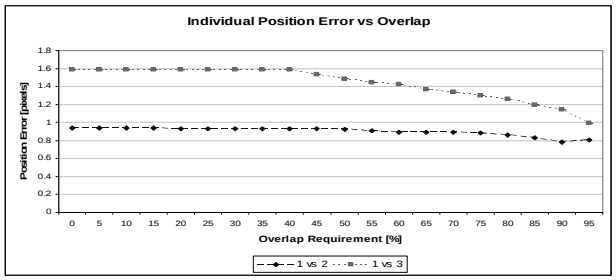

Figure 6. Individual object position accuracy. 1 std dev at 50% overlap are 0.93 and 2.02, respectively.

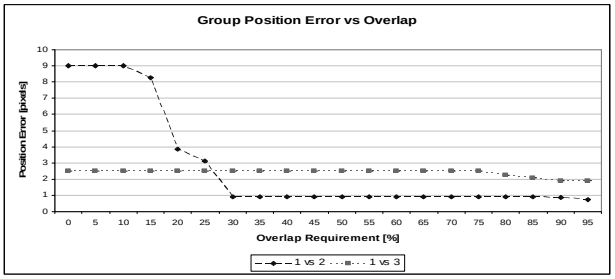

Figure 7. Group position accuracy. 1 std dev at 50% overlap are 0.67 and 2.50, respectively.

In Figure 6 we see the discrepancies of individual objects' centre of box positions, measured in pixels, for Observers 2 and 3. Figure 7 shows the same for group boxes and here we see a much larger difference where Observer 2 is 9 pixels off Observer 1's estimates when not requiring much overlap. This is because Observer 2 sometimes has smaller boxes for groups (see Figure 9), but does agree on one of the corners, thereby causing an offset of the centre of box position. For higher overlap requirements all observers agree very well indeed, within 1 to 3 pixels. This suggests that human marking is quite accurate with regards to position.

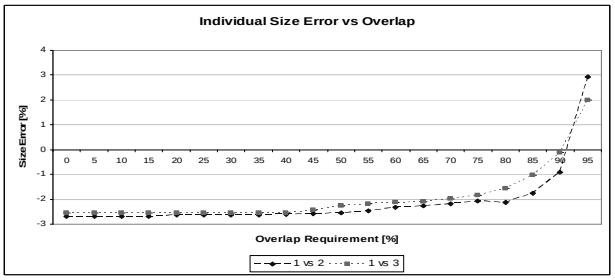

Figure 8. Individual object size accuracy. 1 std dev at 50% overlap are 11.4 and 5.14, respectively.

Figure 8 and Figure 9 show the difference in sizes of bounding boxes as a percent of Observer 1's boxes. Here there is good agreement of all three observers.

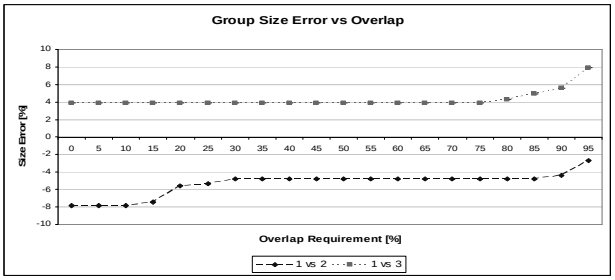

Figure 9. Group size accuracy. 1 std dev at 50% overlap are 3.01 and 11.4, respectively.

For group boxes the sizes are slightly more different and here we see that Observer 2's boxes are generally a bit too small and Observer 3's generally a bit too large, but again all within a reasonable agreement.

To conclude, for spatial tracking the three observers often disagree on which marginally active objects to track, both for individual objects and for groups of objects, with up to 80% additional detections in some cases. For the individual objects, both position and size are in good agreement, usually within 1-2 pixels for position and within 3% for size. One must conclude that this is the case for group positions and sizes as well, where these numbers rise to only 2-3 pixels for positions (for reasonable overlap requirements) and 4-8% for sizes, in spite of groups being a somewhat more subjective entity.

3. Temporal tracking

When considering the temporal aspects of tracking we investigate the discrepancy between when the observers first noticed the object entering the scene, when the observers last saw the object exiting the scene and how frequently the object track was lost in between.

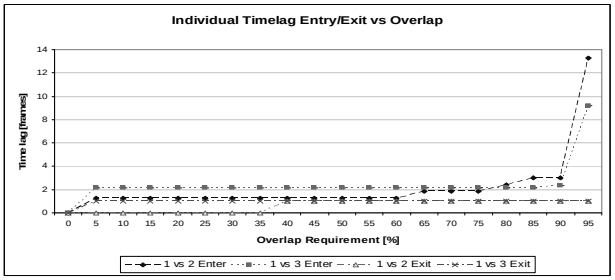

Figure 10. Individual object timelag enter and exit. 1 std dev at 50% overlap are 3.53, 4.83, 0 and 0, respectively.

In Figure 10 we now see four graphs, the average timelags in entering and exiting the scene for both Observer 2 and 3, as compared with Observer 1. We see that for most overlap requirements there is a very good agreement between the three observers (within 2 frames = $2/25^{th}$ of a second). All track 10 targets.

For tracking the object until it exits the scene they all agree perfectly to within 1 frame, which must be said to be a very good result indeed.

For groups in Figure 11 the results show slightly higher timelags. This measures the point in time when the observers determine that two or more individuals now participate in a group activity. Here agreeing within 5 frames ($1/5^{th}$ of a second) is actually a very decent result.

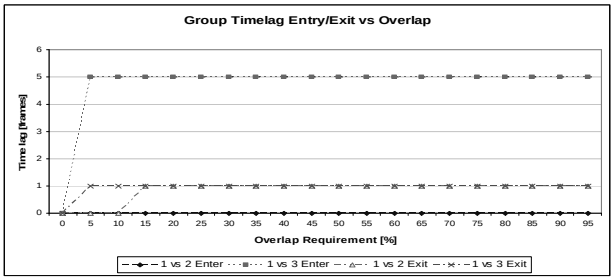

Figure 11. Group timelag enter and exit. 1 std dev at 50% overlap are all 0.00.

The dropped frame rates in Figure 12 and Figure 13 measure in how many frames the tracking of individual objects and groups were lost (or dropped), as compared with the whole period of being visible.

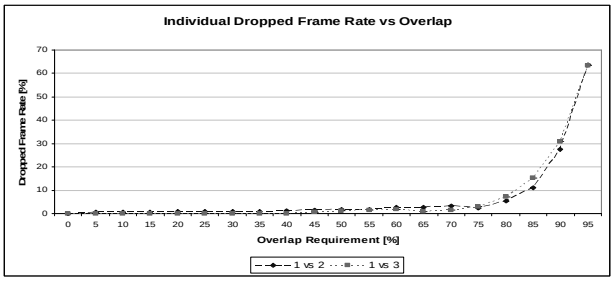

Figure 12. Individual object dropped frame rate. 1 std dev at 50% overlap are 5.03 and 1.89, respectively.

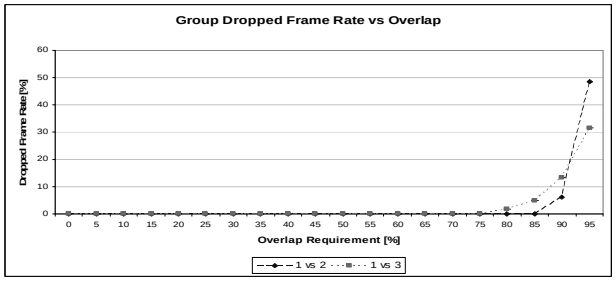

Figure 13. Group dropped frame rate. 1 std dev at 50% overlap are both 0.00.

When requiring an overlap greater than 40% there is a slight wavering up to about 3% dropped frames, which is explained by one object being tracked by Observer 1 even when almost obscured, where Observers 2 and 3 both gave up. When the requirement goes above 75% overlap the observers start to disagree on the position and size enough that many dropped frames are detected.

For groups it is slightly surprising that the observers agree better, as groups are per definition a more difficult entity to agree on.

4. Human behaviour labelling

In addition to the geometric description of tracked individuals and groups, for each target a behaviour hypothesis is assigned which consists of four labels: movement, role, situation and context. These labels describe, respectively, the basic activity level of a target, its role in the situation, the specific situation it is in, and its general behaviour (context). Individual and group targets have different labels, since group behaviours are more general and a group behaviour can be decomposed in different individual behaviours.

The purpose here is to evaluate the agreement of humans in labelling the behaviour of individual and group targets.

4.1 The comparisons done

To evaluate the agreement between people when they assign labels for the targets, 6 comparisons between 3 observers were done and averaged. In these comparisons an intersection of 70% between boxes was used.

For each pair, for example Observer 1 and Observer 2, comparisons of 1 with 2 and 2 with 1 were done. The reason for both directions is that, although the absolute number of instances of one label which agree or disagree for both observers are the same, the percentage of instances which agree on the label will be different, since the total number of instances is different for each observer.

As an example, when comparing the movement for individual targets, for the label “inactive”, Observer 2 has 100.00% agreement with Observer 1, but Observer 1 has only 68.03% agreement with Observer 2 (see Table 1). This is because all 100 instances of the label for Observer 2 agree with Observer 1, but only 100 of the 147 instances for Observer 1 agree with Observer 2.

Averaging the 6 comparisons gives a result unbiased by the direction of comparison.

Inactive	Observer 1	Observer 2
Instances in agreement	100	100
Percentage of agreement	68.03	100.00
Total number of instances	147	100

Table 1. Inactive label comparison for individual boxes for Observers 1 and 2

Note that in the results shown here, only the targets from the first observer, whose boxes overlap more than 70% with the second observer, are used. This is done to avoid mixing the issue of consistency in target selection, seen in the previous section, with the issue of consistency in behaviour labelling.

4.2 Comparison for individual targets

Table 2 shows the average of the comparison of labels for the movement of individual targets. Most labels agree ~80%, but there is not complete agreement by the observers about the movement labels. The main difference is between labels "active" and "walk".

Movements	inactive	active	walking	running	total
inactive	84.55	15.45	---	---	712
active	7.70	32.03	60.07	0.20	1480
walking	---	11.82	86.61	1.57	7444
running	---	---	19.51	80.49	533

Table 2. Average Movement labels comparison for individual boxes

This is mostly due to Observer 1 who labelled a fighting person who was active as 'walking' rather than 'active' (as labelled by Observers 2 & 3).

Table 3 compares individual roles. There is almost complete agreement about which targets are walking, but there is some disagreement about which targets are browsing. This is due to Observer 2 and Observer 3 who marked targets as browsing that were close to a browsing area, but were only passing by.

Roles	browser	walker	leaving victim	leaving group	leaving object	total
browser	47.58	52.42	---	---	---	248
walker	1.31	98.69	---	---	---	9921

Table 3. Role labels comparison for individual boxes[1]

Situations	moving	inactive	browsing	shop enter	shop exit	total
moving	94.43	5.50	0.07	---	---	8440
inactive	30.45	61.17	8.37	---	---	1481
browsing	2.42	50.00	47.58	---	---	248

Table 4. Situation labels comparison for individual boxes

Table 4 shows the comparison between situations. The differences here are much like the movement and role comparisons. In fact, they are derived from them since if an observer uses "inactive" or "walking" for the movement it will use correspondingly "inactive" or "moving" for the situation. Similar reasoning is true for "browsing" and "inactive". The same is true for the context (shown in Table 5). Here the differences are mostly due to the mislabelling of browsing. All targets labelled as "immobile" by Observer 1 were labelled as browsing by the other two observers.

Context	browsing	immobile	walking	drop down	total
browsing	47.60	52.14	0.26	---	1166
immobile	84.63	---	15.37	---	657
walking	---	2.12	97.88	---	4868
drop down	---	---	0.03	99.97	3478

Table 5. Context labels comparison for individual boxes

The differences between the labels for individual boxes show that observers are biased by their own interpretations of what is happening in the scene, and that in some cases those definitions are very different or ambiguous. When unambiguous, the results agree more than 80% and in some cases close to 100%.

4.3 Comparison for group targets

Group behaviours are more consistent than individual behaviours since they are in some sense an average of these, and hence more robust. Nevertheless, some of the same problems that appeared for individual targets appear again here. Table 6 shows the ambiguity between "active" and "movement".

[1] *Some labels were removed from Tables 3, 4, 5 and 7 for layout reasons and because they were not used.*

Movement	inactive	active	movement	total
inactive	---	---	---	---
active	---	91.04	8.96	859
movement	---	63.11	36.89	122

Table 6. Movement labels comparison for group boxes

Situations	fighting	moving	joining	splitting	leaving victim	total
fighting	90.01	---	3.56	---	6.43	731
moving	---	---	---	---	---	---
joining	14.36	---	85.64	---	---	181
splitting	---	---	---	---	100.00	1
inactive	---	---	---	---	---	---
leaving victim	69.12	---	---	1.47	29.41	68

Table 7. Situation labels comparison for group boxes

Table 7 shows that observers may have doubts regarding when some situations begin or end. This is the reason for the differences for the "leaving victim" label.

The tables for roles and context for groups were suppressed because they agree 100% on the role ("fighters") and the context ("fighting"), despite the fact that individual behaviour has lots of disagreement. This may be due to the fact that it may be difficult for a observer to decide if the target is moving or not, but they can identify with great certainty the purpose of the targets as a group.

5. Semantic consistency measurements

In an approach inspired by Martin [7], we also define a consistency measure for the semantic labellings. However, there are some differences between image region segmentation and video behaviour labelling that affect the definition of the consistency measure. While there probably is a notion of 'underlying truth' to the segmentation, which individual labellers approximate, there is probably not a hierarchical 'level-of-detail' in the labelling here (in the sense used by Martin), at least not with the currently defined labelling grammar.

Hence, the notion that two segmentations might be differently refined views of the same underlying structure is not the case here.

A second difference is that here each temporal segment has type labels and we require that the types match. We examine consistency for each category in {movement, role, situation, context}.

A third difference is that tracks start and end at unpredictable frame numbers, and frames where no tracking is present should not count towards the consistency measure. This is unlike image segmentation where every pixel belongs to a segment.

Based on these differences, we define a consistency measure for each category as follows. Let $d_{i,j,p,t}$ be the number of frames where property label for person p for Observer i is different from the property for Observer j (or Observer j did not track the person). This is computed for each category type t. Let $n_{i,p,t}$ be the number of tracked frames for person p for Observer i. Then the consistency measure is:

$$c_{i,j,t} = \frac{\sum\left(d_{i,j,p,t} + d_{j,i,p,t}\right)}{\sum\left(n_{i,p,t} + n_{j,p,t}\right)} \qquad \textit{Equation 1}$$

Using this consistency measure on the three ground truth datasets, we report these consistencies. Here 0 means perfect consistency and 1.0 means completely different labelling. Table 8 shows the raw consistency.

Observers	Movement	Role	Situation	Context
1 & 2	0.5396	0.4665	0.5180	0.5557
1 & 3	0.2862	0.2413	0.2779	0.3551
2 & 3	0.5608	0.4848	0.5530	0.4975

Table 8. Raw labelling consistency.

From these results, we see there is a lot of variation in the symbolic labelling of the test sequences, in all categories. What are the causes of the differences?

1. A small target in a dark part of the background has been omitted by Observer 3. Two other minor targets were separately detected by Observers 1 & 2.
2. One target left a toe in the image when exiting from the scene and it was continued to be tracked by Observer 2.
3. One target was nearly stationary and Observer 2 decided that it was to be tracked and analysed.
4. One person is standing idle near a browsing point, but is not looking at the information. Thus there is ambiguity about whether the target is browsing (Observer 1) or merely idle.
5. There is some disagreement about whether people are running or not.

If we only consider frames where both observers have tracked the person (issues 1,2.3), then the consistency figures (over ~1700 matched target frames each) are:

Observers	Movement	Role	Situation	Context	Frames
1 & 2	0.1683	0.0362	0.1293	0.1973	1794
1 & 3	0.0918	0.0347	0.0812	0.1794	1700
2 & 3	0.1504	0.0034	0.1353	0.0279	1789

Table 9. Labelling consistency where both observers have tracked the person.

This shows that labelling consistency is much better than the initial figures suggest. What is reflected in Table 9 is now mainly a difference in interpretation, rather than a difference in target appearance or timing.

If we now consider issues of semantic ambiguity (issues 4,5), then more consistency is observable. In particular, labelling people as running or walking can be subjective. Similarly, people that are not walking can be treated as either active or inactive, depending on how much they move. Situations and contexts can also be labelled as either idleness (loitering) or browsing, depending on the observer's judgment about the person's intent and gaze direction.

We retested the consistency, allowing all of the ambiguities listed above, and also declared consistency if the labels were consistent within plus-or-minus 20 frames (to allow some differences in timing - see below).
With this we get these statistics:

Observers	Movement	Role	Situation	Context
1 & 2	0.0675	0.0296	0.0675	0.0265
1 & 3	0.0178	0.0281	0.0178	0
2 & 3	0.1065	0	0.1065	0.0272

Table 10. Labelling consistency where both observers have tracked the person - retested.

Thus, there is much better consistency, and it suggests only about maximum of about 11% labelling errors, and it could be as low as about 3%.

We also evaluate the timing on the starting and ending of segments that match all of {movement, role, situation, context}. These were selected over all possible pairs (1 & 2, 2 & 3, 1 & 3). Because there are possible labelling errors, we use a robust statistics estimate of the mean and standard deviation based on the median. This gives us (over 28 samples):

Item	Mean	Std. Dev.
Start frame difference	0	11.9
End frame difference	-1	20.8

Table 11. Consistency of starting and ending segments.

This suggests that observers are, on average, choosing more or less the same times to start and end actions, but there is about a ±1-2 seconds variation (at 25 fps) in deciding when the events occur.

6. Conclusion

We have investigated how well three human observers agree on spatial, temporal and behavioural observations in a standard video sequence. We found that the observers disagree significantly on which objects and groups of objects to track (participants) and which to leave out (non-participants), with up to 80% extra objects in one case. When considering the accuracy with regards to object and group positions and sizes the three observers agree very well, within 1-2 pixels on object positions, 1-3 pixels on group positions, 3% on object sizes and 4-8% on group sizes.

For the temporal tracking results, we investigated how quickly the observers noticed the objects and groups after their initial appearance, how well they agreed on the time of exit, and how often during the track they lost track of the objects and groups. We found that for object appearance the observers agreed within 2 frames ($2/25^{th}$ of a second) and for groups within

5 frames ($1/5^{th}$ of a second). For exiting they agreed within 1 frame on both individual objects and groups. The lost targets were less than 3% for individuals and a solid 0% for groups.

When considering the semantic labelling, there is much more difference between the observers, but this is explained by four factors:

1. Observers are biased by their own interpretations of what is happening in the scene.
2. There is a possibility for ambiguous interpretations of the target's behaviour, in particular between whether the person is running or walking, immobile (loitering) or browsing, active or inactive.
3. There are slight variations in timing of when behaviour changes. There seems to be a standard deviation of 0.65 second about a zero mean for detecting changes.
4. There may be some targets that an observer chooses to not mark, because the target hardly moves, or the target is small, distant and in a low contrast area (i.e. the marker may not have observed the target).

Allowing these ambiguities gives consistency at the 11% level. This seems to set a limit on the quality of semantic ground truth for video sequence data. Therefore, an automatic interpretation program that allows these ambiguities will achieve a better performance, but is unlikely to achieve less than 11% semantic error as humans do not always even agree, particularly on movement levels.

Our consistency evaluation has several problems that are hard to fix:

1. The sample size is small: 3 labellings of the same sequence is not enough to give reliable statistics.
2. The three observers were PhD or undergraduate students in our laboratory and thus were biased by the discussions of the sorts of recognition approaches we were considering, and by having shared supervision. On the other hand, the video sequence used was one of the first that we labelled, and so some of the confusions the researchers were working through are similar to the issues that arise in the labelling by new people. Thus the results are not totally unrealistic.

We believe that this is the first attempt to assess video sequence ground truth labelling from both a detection and behaviour analysis perspective. Hence, the results presented here are usable as a baseline until a larger experiment is done.

Acknowledgments

This research was supported by the CAVIAR project, funded by the EC's Information Society Technology's programme project IST 2001 37540.

References

[1] S. Crawford-Hines, “Learning from the Expert: Improving Boundary Definitions in Biomedical Imagery”, *Knowledge-Based Intelligent Information and Engineering Systems, KES 2003*, 7th International Conference, Oxford, UK, September 3-5, 2003, pp 653-659, 2003.

[2] J. L. Crowley, J. Coutaz, G. Rey and P. Reignier, "Perceptual Components for Context Aware Computing", *UBICOMP 2002, International Conference on Ubiquitous Computing*, Goteborg, Sweden, September 2002.

[3] D. Doermann and D. Mihalcik, “Tools and Techniques for Video Performances Evaluation”, *International Conference on Pattern Recognition, ICPR00*, pp 167-170, 2000.

[4] R. B. Fisher, “The PETS04 Surveillance Ground-Truth Data Sets”, Proc. Sixth *IEEE Int. Work. on Perf. Eval. of Tracking and Surveillance (PETS04)*, pp 1--5, Prague, May 2004.

[5] T. List and R. B. Fisher, “Computer Vision Markup Language”, Proceedings of *International Conference on Pattern Recognition, ICPR04*, Cambridge, Vol 1, pp 789-792, 2004.

[6] V.Y. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, D. Doermann and T. Drayer, “Performance Evaluation of Object Detection Algorithms”, *International Conference on Pattern Recognition, ICPR02*, pp 965-969, 2002.

[7] D. R. Martin, “An Empirical Approach to Grouping and Segmentation”, PhD Thesis, Section 2.2, University of California, Berkeley, 2002.

[8] D. Martin, C. Fowlkes and J. Malik, "Learning to Detect Natural Image Boundaries Using Local Brightness, Color and Texture Cues", IEEE Trans. on Pattern Analysis and Machine Intelligence, 26(5), pp 530-549, May 2004