Event-Based Tracking Evaluation Metric

D. Roth[1], E. Koller-Meier[1], D. Rowe[2], T.B. Moeslund[3], L. Van Gool[1]

[1] Computer Vision Laboratory, ETH Zurich, Switzerland
{*droth,ebmeier,vangool*}*@vision.ee.ethz.ch*

[2] Computer Vision Center / UAB, Barcelona, Spain
drowe@cvc.uab.es

[3] Laboratory of Computer Vision and Media Technology, Aalborg University, Denmark
tbm@cvmt.dk

Abstract

This paper describes a novel tracking performance evaluation metric based on the successful detection of events, rather than low-level image processing criteria. A general event metric is defined to measure whether the agents and actions in the scene given by the ground truth were correctly tracked by comparing two event lists using dynamic programming. This metric is suitable to evaluate and compare different tracking approaches where the underlying algorithm may be completely different.

Furthermore, we introduce an automatic extraction of those semantically high level events from different types of low level tracking data and human annotated ground truth. A case study with two different trackers on public datasets shows the effectiveness of this evaluation scheme.

1. Introduction

Performance evaluation of multi-object trackers for surveillance has recently received significant attention. Apart from functional testing, there are several additional reasons for evaluation: measuring improvements during development, scientific justification, benchmarking with competitors, measuring the progress in the whole research community or commercial and legal purposes. Evaluation metrics for surveillance are almost as numerous as multi-object tracking methods themselves [1]. A problem is that they mostly address issues which do not directly tie in with the overall semantic interpretation of the scene that users would be most interested in. As an example, assessments of pixel-precise target detections are relevant for the evaluation of subcomponents like figure-ground segmentation, but fall short of determining whether a system can make sense of what is going on in the scene. The interpretation of scores in terms of what they convey about the correct or incorrect analysis of actions is often difficult. As a result, human visual inspection is still needed to compare and estimate general performance and limitations.

Our novel tracking performance evaluation method is motivated by the fact that humans conceptualize the world in terms of events and objects, and our metric aims to imitate this behavior by evaluating tracking performance on such higher conceptual level. Instead of comparing trackers and ground truth data directly on a low semantic level, we extract different types of higher level events such as *entering the scene*, *occlusion* or *picking-up a bag* from the available data. Our metric then focuses on the completeness of such event detection to do the evaluation.

This work is part of the research project HERMES [2], which aims at developing an artificial cognitive system allowing both recognition and description of a particular set of semantically meaningful human behaviors from videos. The system to be developed will combine active and passive sensors. HERMES pays attention to three resolution levels of scene analysis. Depending on distance, people's actions can be analyzed as moving blobs, or as articulated body gestures, or on the basis of facial expressions. This research project has set out to interpret and combine the knowledge inferred from those 3 different motion categories. An important aspect is the combination of low level vision tasks with higher level reasoning, as well as the evaluation of different visual systems which cover one or multiple semantic levels. The recognized behaviors will then be used for natural language text generation and visualization. Our event-based tracking evaluation metric can therefore also be seen as a first attempt towards a unified metric which could combine different visual tracking disciplines on a higher conceptual level.

Our proposed method comprises the following advantages:

- The lengths of trajectories do not influence the metric making it independent of the frame rate and density of the ground truth labeling.
- The type of events taken into account for the final metric can be fine tuned for different application scenarios.
- Easy integration into higher level event and object detection frameworks.
- The metric directly helps to improve tracking algorithms by identifying:
  - difficult trajectories
  - difficult scene locations
  - difficult situations
  - difficult event types
- Fast generation of ground-truth data as not every frame needs to be annotated in full detail, as long as the events can be reliably extracted from sparse annotation.
- Reuse of already available ground truth data by automatic conversion into our novel event-based representation.
- Minimizing the human factor within the ground truth data and its influence onto the metric by means of event-based evaluation on a higher level.
- A precise distance measurement between objects in real-world coordinates eliminates the need to define unreliable 2D bounding box distances.
- Aims at minimizing the need for human visual inspection of results, allowing faster testing of new algorithms or longer sequences.
- Establishing a least common denominator to represent tracking data which is versatile to handle many different output formats.

1.1. Previous Work

Related work in the field of performance evaluation for object tracking can roughly be classified into different semantic classes that specifically address one or more of the following semantic levels:

- pixel-level [3, 4]
- frame-level [5, 3]
- object trajectory level [5, 6]
- behaviors or higher level events [7]

Desurmont *et al.* [7] which is most closely to our work presented a general performance evaluation metric for frequent 'high level' events where they use dynamic programming to specifically address the problem of automatic realignment between results and ground truth. Their definition of an event is however very much limited to the detection of blobs crossing predefined lines in order to count people passing by. The same limitation applies also to the use of dynamic programming, whereas our method integrates additional location information to the matching problem. Bashir and Porikli [5] presented a set of unbiased metrics on the frame and object level which leaves the final evaluation to the community. However, the total number of 48 different metrics make the interpretation difficult. Aguilera *et al.* [3] presented an evaluation method for the frame and pixel level, all based on segmentation. The pixel-level metric is currently used for the online service called PETS metrics [8].

Wu *et al.* [6] evaluate their body part detector-based tracker using five criteria on the trajectory level which cover most of the typical errors they observed. Furthermore, occlusion events were separately evaluated defining short and long scene or object occlusions. The metric then gives the number of successful handled occlusions against all occlusions of a certain category by dataset.

The ETISEO Project [9] aims at acquiring precise knowledge of vision algorithms by inviting multiple different institutes to report on a general corpus of video sequences. Different metrics were proposed and a final evaluation is still in progress. Within ETISEO Nghiem *et al.* [4] presented an interesting approach where they evaluate multiple trackers on isolated video processing problems of different difficulties.

The rest of the paper is organized as follows: Section 2 defines the event metric; Section 3 introduces the case study and their relevant events. Section 4 shows the experiment results, and Section 5 discusses our findings and draws a conclusion.

2 Event-Based Tracking Metric

This Section explains how our proposed event-based tracking metric is defined, generated and evaluated. Figure 1 shows an overview of the application of the event metric. Event information is both extracted from the ground truth data and from the tracker. The lists can then be compared by using the proposed event metric definition resulting in a score that reflects how well the tracker is able to handle the specific sequence.

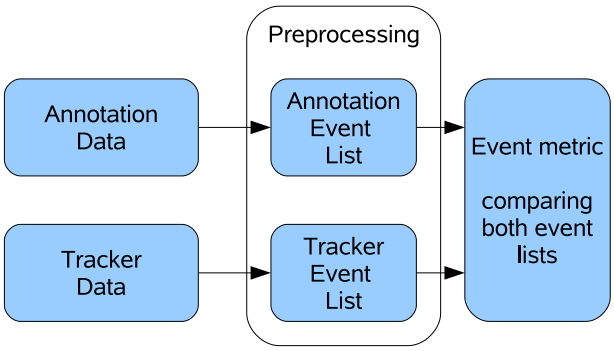

Figure 1: Evaluation scheme

2.1 Event concept

The event is our basic building block of the evaluation metric. We represent our higher semantic level as a list of events which together describe the action in a scene similar to the conceptual level of humans and similar to human language. Each event describes an action conducted by an individual similar to subjects and verbs in natural language. Gerber and Nagel [10, 11] describe a system which extracts 'occurrences' from road traffic scenes. Those 'occurrences' have a broader meaning than our 'event' which is limited to actions without a duration. Our 'event' always describes instant actions happening at one particular point in time unlike the 'occurrences' which distinguish explicitly between *perpetuative*, *mutative* and *terminative* actions.

This simplification has several advantages for the evaluation task because continuous unchanged states are extraneous. Furthermore, the evaluation metric can be simplified if all events are instantaneous. Finally, the use of events does not prevent us from modeling perpetuative actions as we can enclose them with starting and ending events as well as separating mutative evolving actions accordingly. Other definitions of 'events' can be found in the literature [1], but none of these definitions are general enough to suit our needs.

An event $\boldsymbol{E}$ is a 4-tuple and always consists of an event-type $\mathcal{P}$, a point in time $\mathcal{T}$, a location $\mathcal{L}$ and it is related to one object $\mathcal{O}$. The event-type $\mathcal{P}$ identifies an action or interesting change in the scene as described in the next subsection. $\mathcal{T}$ is measured in seconds and fraction thereof, computed from the frame rate and frame number when an event occurs in the sequence. The event location $\mathcal{L}$ is defined as the 3D base point in world coordinates of the associated object. Given the camera calibration and a ground plane assumption a 2D to 3D correspondence is given in most cases. Objects $\mathcal{O}$ are numbered with unique IDs.

2.2 Event Types

Event types $\mathcal{P}$ are selected in such a way that they are relevant for the application, have higher level meaning, can unambiguously be identified from ground truth as well as from the tracker results and they need to be atomic. In order to simplify the final evaluation metric described in Section 2.4 we treat each $\mathcal{P}$ individually which leads to the following examples of event types that are used for the case study in Section 3:

- Entering / Leaving scene
- Starting / Ending occlusion
- Entering / Leaving a specific area like a shop or a pedestrian crossing

Many more events could be considered for other applications, for example:

- Pointing at something / Being pointed at
- Starting / Ending *walking, running, standing*
- Picking up a bag

Actions with a certain duration such as the presence in the scene, movement attributes like running or an occlusion are split into a starting and ending event. Very short actions such as picking up something are single events. Interactions are handled implicitly by expecting the same event from each involved actor, like for occlusions. In case of directed actions such as pointing, two different event-types are used. Furthermore, it is possible to define special areas in the scene such as a pedestrian crossing or waiting areas which can trigger events.

Within the scope of this paper, we do not further expand or group events into different hierarchical layers even though some events would have a semantic relationship. Within such a hierarchical event-logic, an 'entering' and a 'leaving' event for the same object could be grouped into a 'object was seen' event for example. This would further require to define a distance measure between events in order to correctly evaluate partial correct event-structures. Defining semantic distances is very difficult especially for the general task of multi-object tracking. However, for specific applications and well constrained tasks a hierarchical event-logic could indeed be defined as shown in [10, 11].

2.3 Event Generation

In order to generate events from either manually labeled data or continuous tracker output, only the four basic building blocks of an event $\boldsymbol{E}(\mathcal{P}, \mathcal{T}, \mathcal{L}, \mathcal{O})$ need to be extracted from the data. This allows to compare different types of trackers with different flavors of annotation data. However, individual conversion methods have to be used to generate the events depending on the type of the underlying data.

$\mathcal{T}$ is measured in seconds and $\mathcal{O}$ can directly be extracted from almost any data type. $\mathcal{L}$ is more difficult to extract as it most often needs further processing to find the 3D base point from 2D image coordinates of an object given the camera calibration. The most difficult part is the definition and extraction of events, which will be discussed in more detail.

Generally, event extraction rules can be defined as complex as desired, using additional data such as camera calibrations or location maps, which are not necessarily required during annotation or tracking. Using the same formalism and categorization of Gerber and Nagel [10, 11], we can define an event by its pre-condition which has to be satisfied *before* the event is happening and a post-condition once the event has happened. Section 3.3 gives concrete examples.

In order to reliably extract events, it is often important to take more information into account than just $\mathcal{P}$, $\mathcal{T}$, $\mathcal{L}$, $\mathcal{O}$. The additional information required to reliably define an event has to be chosen carefully in order to allow comparison between different trackers and sequences.

2.4 Evaluation Metric

Our evaluation metric is based on comparing the list of events extracted from ground truth data and the list of events extracted from the trajectories generated by the tracking algorithm. The evaluation can best be described as a pipeline of 3 steps where we first use dynamic programming to find the best matches between events of the same $\mathcal{P}$ from both lists. Then we compute the different measurements based on the matched events. Finally, we analyze the event matching for each object individually in order to measure changing object identities. Due to the characteristic of the 4-tuple of each event the metric can exploit either binary matches or continuous distances as shown in Table 1.

$\boldsymbol{E}$	building blocks	binary match	cont. distance
$\mathcal{P}$	event type	X	-
$\mathcal{T}$	time	thresholded	X
$\mathcal{L}$	location	thresholded	X
$\mathcal{O}$	object ID	X	-

Table 1: Evaluation metric

As described in Section 2.2, we do not define a distance metric between different event-types. We only match and evaluate each event $\mathcal{P}$ separately.

2.5 Evaluation Pipeline

In the first step of our evaluation pipeline we use dynamic programming techniques to match events from the same event type. As distance measurement between two events i, j we combine $\mathcal{T}$ and $\mathcal{L}$ into one distance in the following manner:

$$dist_{i,j} = min(\alpha|\mathcal{T}_i - \mathcal{T}_j| + ||\mathcal{L}_i - \mathcal{L}_j||, maxdist)$$

Where $||...||$ is the Euclidean distance, α is a scaling factor in order to allow to compare the different units of seconds and meters. The parameter *maxdist* is a maximal distance above which the match is considered to have failed. For our experiments, we have chosen α in such a way that a difference of 5 seconds respectively 12 meters is equal *maxdist*. These values are chosen with a generous margin above $\mathcal{T}_{ave}$ and $\mathcal{L}_{ave}$ as we do not want to penalize reasonable time and location deviation at this early stage in the evaluation pipeline. The dynamic programming will give us the optimal match between ground truth and tracker events according to the above time and location distance as shown in Figure 2. All matches which stay below the maximal distance are counted as true positives (TP). Events on the ground truth list which have no corresponding event in the tracker list or match with *maxdist* are counted as false negative (FN). Events from the tracker which could not be matched to any ground truth event are counted as false positive (FP). This first evaluation measures how well the defined event types were correctly handled by the tracker regardless of a correct and continued identification of the subjects in the scene.

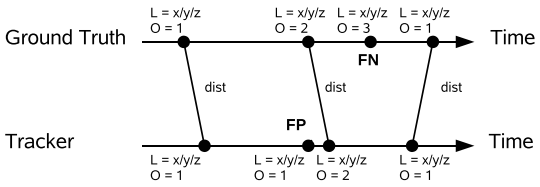

Figure 2: Event matching of same type

In a second step, we compute the average time $\mathcal{T}_{ave}$ and location $\mathcal{L}_{ave}$ deviations from ground truth of all correct matches (TP) to measure the accuracy of the tracker. FP and FN are important to further identify time periods and locations where errors in the tracking occurred.

In a third step, we take a closer look at all ground truth subjects with a TP *Entering Scene* event and their corresponding $\mathcal{O}$. This metric answers the question: how many of all events were detected correctly of an object. We compute the percentage of correct events TP out of all events. This percentage is a measure for the tracking quality of a certain object. Furthermore, we count the total number $\mathcal{O}_{tot}$ of different $\mathcal{O}$ for the ground truth object it was matched to, in order to measure identity changes.

Events in the first and last frame of the ground truth are not evaluated. Tracker events which can be matched

to those events are discarded. This prevents the evaluation of events, which might have happened before the first frame and obviously could not be annotated correctly such as people already present in the first frame.

3 Case Study

In order to verify the versatility of our novel evaluation metric we conduct a case study where we apply two different tracking algorithms to two well-known benchmarking data sets: CAVIAR and PETS2001. In this Section we first present the two data sets and analyze them in order to identify the relevant events. These are then described using the notation described above. Furthermore, we briefly present the two trackers to test our evaluation scheme.

3.1 CAVIAR Data Set

From the numerous CAVIAR datasets [12] we choose the short *OneLeaveShopReenter1cor* sequence to demonstrate the event extraction by evaluating six different event types, relevant for this sequence: *Entering scene*, *Leaving scene*, *Start occlusion*, *End occlusion*, *Entering shop* and *Leaving shop*. The ground truth events were extracted from the public available XML annotation data. The shop area shown in Figure 3 is the sole additional information to this evaluation. The base point of the objects in world coordinates are calculated with the camera calibration as given on the CAVIAR web-site.

Figure 3: CAVIAR OneLeaveShopReenter1cor sequence

3.2 PETS 2001 Data Set

As another example we experimented with the well known PETS2001 dataset [13] as it consist of additional challenges for our metric. Fine tuning of the automatic event extraction was needed due to the special XML format and the different object types such as cars and people. For this evaluation we ran the trackers for the first 1570 frames. Relevant events for this sequence are *Entering scene*, *Leaving scene*, *Start occlusion* and *End occlusion* which were extracted automatically.

3.3 Event Description

Even though the event types are semantically clearly defined, we briefly describe their actual implementation for the different data formats. *Entering and Leaving scene* events are found in those frames where an object is seen the first and last time in the sequence. Due to common instabilities in size and localization during entering and leaving, we apply a weighted average filter to the object position. The object size given by the bounding box area is used as a weighting factor for $\mathcal{L}$ over 10 frames. Human annotations tend to contain single hands, arms or heads of persons while they appear or disappear, which would give very wrong base point assumptions if not filtered. *Start and End occlusion* events are triggered as soon as two objects overlap. While the initial start event needs a significant overlap in order to filter out small contacts.

Special area events such as *Entering and Leaving Shop* use an additional marked image area for which object movements into or out of the area trigger such events. Again, we filter the number of events for each object to prevent event bursts due to tracking location inaccuracies.

3.4 Tracker 1

The architecture of the first tacker [14] is based on a modular and hierarchically-organized system (see Figure 4). A set of co-operating modules, which work following both bottom-up and top-down paradigms, are distributed through three levels. Each level is devoted to one of the main different tasks to be performed: Target Detection, Low-Level Tracking (LLT), and High-Level Tracking (HLT). Since high-level analysis of motion is a critical issue, a principled management system is embedded to control the switching among different operation modes, namely *motion-based tracking* and *appearance-based tracking*. Tracker 1 has the ability to track numerous targets during grouping and splitting while maximizing the discrimination between the targets and potential distracters.

The system works as a stand-alone application and is designed for offline processing of sequences without the need of real-time operation. For our case study tracking results were exchanged in a proprietary format containing ellipse target representations on a frame by frame basis. To ease further processing of this format, rectangular bounding boxes were computed around the ellipses prior to event extraction, which resulted in similar data over all trackers and ground truth.

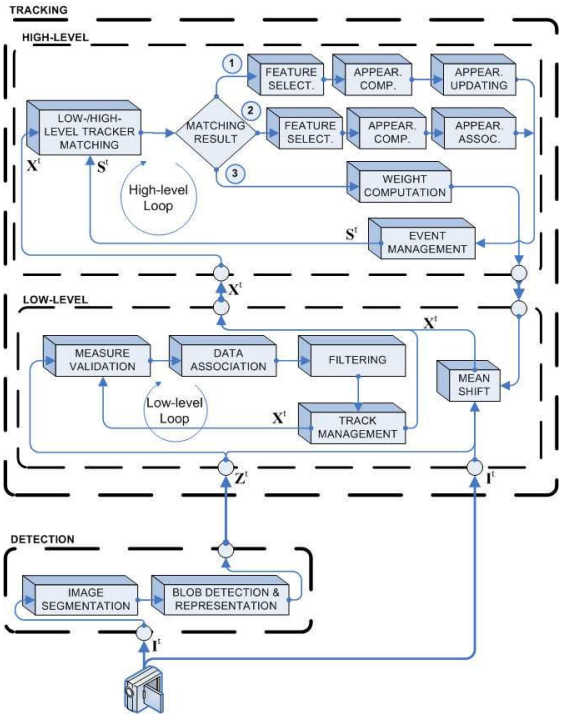

Figure 4: Hierarchical multiple-target tracking architecture

3.5 Tracker 2

The second tracker [15] which we use to test the evaluation metric is a real-time tracker based on segmentation by exploiting a static background. It performs a per-pixel classification to assign every pixel to one of the different objects that have been identified, including a background. The classification is based on the probability that a given pixel belongs to one of the objects given its specific color and position. The object probabilities are determined on the basis of two components. On the one hand, the appearance of the different objects is learned and updated, which yields indications of how compatible observed pixel colors are with these models. On the other hand, a motion model makes predictions of where to expect the different objects, based on their previous position. The appearance models of tracker 2 use Gaussian mixtures in RGB color space with a single Gaussian per-pixel for the background and multiple Gaussians for foreground models. The approach is akin to similar Bayesian filtering approaches, but has been skimmed down to strike a good balance between robustness and speed. Figure 5 sketches the tracking framework with its probability images. Tracking results are generated in the same XML format like CAVIAR containing rectangular bounding boxes for each object.

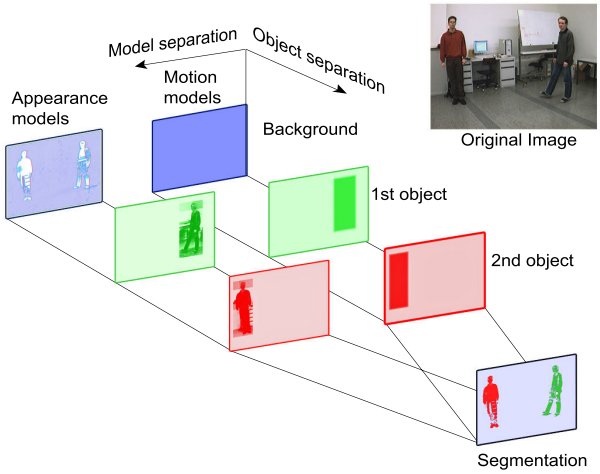

Figure 5: Overview of the Bayesian per-pixel classification tracker

4 Results

This section presents results from our case study where we apply the automatic event extraction on different public datasets to evaluate and compare two different tracking algorithms. Finally, some effects of the metric are discussed.

To test the versatility of our method for different types of ground truth data, we extracted our event representation from a PETS and a CAVIAR sequence. For this process the public available annotation data in different types of XML formats were processed. No additional human ground truth labeling was needed to extract the events automatically.

$\mathcal{P}$	ground truth	tracker 1	tracker 2
Entering Scene	2	2	2
Leaving Scene	1	1	1
Starting Occlusion	2	2	2
Ending Occlusion	2	2	2
Entering Shop	1	1	1
Leaving Shop	1	1	1

Table 2: Detected events of the CAVIAR sequence

4.1 CAVIAR

Table 2 shows the total number of event detections for the *OneLeaveShopReenter1cor* CAVIAR sequence and Table 3 contains the evaluation metric applied to the two object trackers. As can be seen this sequence is tracked perfectly by both trackers. The object-based evaluation is not shown here, as it does not contain more relevant information. Slight differences can be seen in the location accuracy

$\mathcal{L}_{ave}$ where tracker 2 is less accurate than tracker1 due to strong reflections on the floor especially near the shop area. However, no significant time delay can be measured on either tracker.

$\mathcal{P}$ Tracker 1	TP	FN	FP	$\mathcal{T}_{ave}$	$\mathcal{L}_{ave}$
Entering Scene	2	0	0	0.18s	0.17m
Leaving Scene	1	0	0	0.16s	0.81m
Starting Occlusion	2	0	0	0.12s	0.34m
Ending Occlusion	2	0	0	0.16s	0.59m
Entering Shop	1	0	0	0.04s	0.47m
Leaving Shop	1	0	0	0.12s	0.30m

$\mathcal{P}$ Tracker 2	TP	FN	FP	$\mathcal{T}_{ave}$	$\mathcal{L}_{ave}$
Entering Scene	2	0	0	0.32s	1.91m
Leaving Scene	1	0	0	0.08s	0.99m
Starting Occlusion	2	0	0	0.04s	0.12m
Ending Occlusion	2	0	0	0.04s	0.14m
Entering Shop	1	0	0	0.00s	0.81m
Leaving Shop	1	0	0	0.04s	1.44m

Table 3: Event-based evaluation of the CAVIAR sequence

4.2 PETS2001

Table 4 and Figure 4 show for the PETS 2001 DS1 sequence the total number of the event detections. Tables 5, 6, 7 give the evaluation metric applied to the two object trackers. It can clearly be seen that this sequence is more challenging giving more false negatives (FN) and false positives (FP). The evaluation also shows that the offline tracker 1 performs better than the real-time tracker 2. Especially entering and leaving objects are better handled due to the more sophisticated architecture combining bottom-up and top-down paradigms. A closer look onto the failures of tracker 1 shows that most FN and FP are caused by tacking multiple objects as one single object instead of the annotated individuals in the ground truth resulting in several missed occlusion events. Tables 6 and 7 show good results if we just evaluate those ground truth objects which have an associated tracker object. The analysis of tracker 2 shows a high FP numbers of entering and leaving scene events, which is caused by lost tracks during occlusion. This can also directly be seen in Table 7 by the high number of $\mathcal{O}_{tot}$ which counts identity switches.

4.3 Metric

Besides the promising results we also found some limitations of the proposed metric. Given the annotated ground truth, the difficulty of a sequence can not fully be determined from the number, types and density of events. However, the number, types and relationships between the events

$\mathcal{P}$	ground truth	tracker1	tracker2
Entering Scene	10	11	18
Leaving Scene	3	5	15
Starting Occlusion	36	32	28
Ending Occlusion	38	28	26

Table 4: Detected events for the PETS sequence

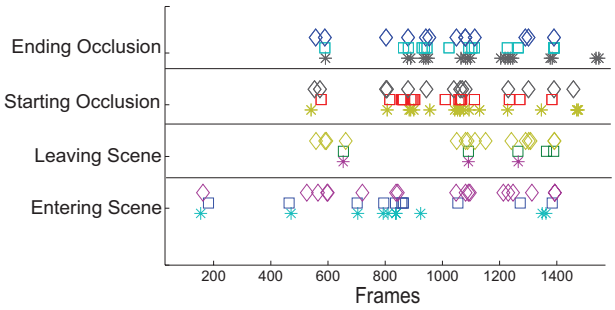

Figure 6: Frames/Event plot for the PETS sequence. Stars are ground truth events, squares from tracker 1 and diamonds show events from tracker 2.

$\mathcal{P}$ Tracker 1	TP	FN	FP	$\mathcal{T}_{ave}$	$\mathcal{L}_{ave}$
Entering Scene	9	1	2	1.28s	0.11m
Leaving Scene	3	0	2	0.02s	0.06m
Starting Occlusion	23	13	9	0.82s	0.28m
Ending Occlusion	20	18	8	0.43s	0.04m

$\mathcal{P}$ Tracker 2	TP	FN	FP	$\mathcal{T}_{ave}$	$\mathcal{L}_{ave}$
Entering Scene	7	3	11	1.8s	0.18m
Leaving Scene	3	0	12	0.7s	1.25m
Starting Occlusion	18	18	10	0.52s	0.29m
Ending Occlusion	16	22	10	1.00s	1.05m

Table 5: Event-based evaluation of the PETS sequence

$\mathcal{O}$ Tracker 1	TP percentage	$\mathcal{O}_{tot}$
GT Object 0	4/4	1
GT Object 1	8/16	3
GT Object 2	12/13	2
GT Object 3	7/11	5
GT Object 5	6/12	2
GT Object 6	3/3	1
GT Object 7	6/10	2
GT Object 8	3/4	2
GT Object 9	2/4	1
Total	51/77 (66%)	

Table 6: Object-based evaluation of tracker 1 (PETS)

$\mathcal{O}$ Tracker 2	TP percentage	$\mathcal{O}_{tot}$
GT Object 0	4/4	3
GT Object 1	10/16	4
GT Object 2	10/13	5
GT Object 3	5/11	6
GT Object 6	3/3	2
GT Object 7	4/10	4
GT Object 9	1/4	1
Total	37/77 (42%)	

Table 7: Object-based evaluation of tracker 2 (PETS)

can give a first estimate for the difficulty of a sequence for a certain event type. Due to the absence of relevant difficult 'events' such as illumination changes or scene occlusions within ground truth and tracking results several difficulties are not directly visible to our metric. Only human inspection of the frames were many failures occur might show the illumination change.

5 Summary and Conclusions

A novel tracking evaluation metric on a semantically higher level was introduced based on events for multi-object tracking. Two different public datasets were automatically processed. They showed the versatility of the metric, which allows to define individual type of events for different application scenarios. Already available annotated ground truth data targeting lower level metrics could be reused and automatically converted into our novel event-based representation. The metric aims at emulating a human visual inspection by conceptualizing the evaluation similar to the human terms of objects and events. This minimizes the need for human visual inspection, allowing faster testing of new algorithms or longer sequences.

5.1 Outlook

Further tests have to show whether the same type of events and metric could be used to also integrate visual detection application such as facial emotion detection or biometric identification.

Acknowledgments

The authors gratefully acknowledge support by the ETH Zurich project blue-c-II, Swiss SNF NCCR project IM2, and EU project HERMES (FP6-027110). Furthermore, we would like to thank Prof. Hans-Hellmut Nagel from the Universität Karlsruhe for his valuable input.

References

[1] T. B. Moeslund, A. Hilton, and V. Kruger, A survey of advances in vision-based human motion capture and analysis, in *Computer Vision and Image Understanding* Vol. 104, pp. 90–126, 2006.

[2] http://www.hermes-project.eu, Hermes website.

[3] J. Aguilera, H. Wildernauer, M. Kampel, M. Borg, D. Thirde, and J. Ferryman, Evaluation of motion segmentation quality for aircraft activity surveillance, in *IEEE Int. Workshop on VS-PETS*, pp. 293– 300, 2005.

[4] A. Nghiem, F. Bremond, M. Thonnat, and R. Ma, A new evaluation approach for video processing algorithms, in *IEEE Workshop on Motion and Video Computing, 2007. WMVC '07.*, pp. 15–15, 2007.

[5] F. Bashir and F. Porikli, Performance evaluation of object detection and tracking systems, in *IEEE International Workshop on PETS* Vol. 5, pp. 7–14, 2006.

[6] B. Wu and R. Nevatia, Tracking of multiple, partially occluded humans based on static body part detection, in *CVPR*, pp. 951–958, 2006.

[7] X. Desurmont, R. Sebbe, F. Martin, C. Machy, and J.-F. Delaigle, Performance evaluation of frequent events detection systems, in *IEEE Int. Workshop on PETS*, 2006.

[8] D. Young and J. Ferryman, Pets metrics: On-line performance evaluation service, in *Proc. 2nd Joint IEEE Int. Workshop on VS-PETS*, pp. 15–16, 2005.

[9] http://www.silogic.fr/etiseo, Etiseo: Video understanding evaluation.

[10] R. Gerber and H.-H. Nagel, Occurrence extraction from image sequences of road traffic scenes, in *Workshop on Cognitive Vision*, pp. 1–8, 2002.

[11] R. Gerber and H.-H. Nagel, Representation of occurrences for road vehicle traffic, in *Artificial Intelligence*, 2007, Article in press, available online.

[12] http://homepages.inf.ed.ac.uk/rbf/CAVIAR, Caviar: Project website, datasets and annotated ground truth.

[13] http://www.cvg.cs.rdg.ac.uk/PETS2001, Pets 2001: Dataset and annotated ground truth.

[14] D. Rowe, I. Reid, J. Gonzlez, and J. Villanueva, Unconstrained Multiple-people Tracking, in *28th DAGM, Berlin, Germany*, pp. 505–514, Springer LNCS, 2006.

[15] D. Roth, P. Doubek, and L. Van Gool, Bayesian pixel classification for human tracking, in *MOTION*, pp. 78–83, 2005.