A Hierarchical Architecture to Multiple Target Tracking

D. Rowe*, I. Huerta*, M. Pedersoli*, S. Guinea+, J. Gonzàlez† , J. J. Villanueva*

* *Computer Vision Centre, UAB, 08193 Barcelona, Spain*
+ *Clínica Plató, 08006 Barcelona, Spain*
† *Institut de Robòtica i Informàtica Industrial, UPC-CSIC, 08028 Barcelona, Spain*
E-mail: drowe@cvc.uab.es

Abstract This work presents an architecture based on a modular and hierarchically-organised system to perform Multiple-Target Tracking (MTT). 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 a main task: Target Detection, Low-Level Tracking (LLT), and High-Level Tracking (HLT). A principled event management is embedded in the system. The system is allowed to switch among Motion-based Tracking (MBT) and Appearance-based Tracking (ABT).

Keywords: Multiple-target tracking; Trajectory analysis; Event management; Appearance-based tracking.

1 Introduction

Human beings, as well as a great diversity of animal species, have developed an amazing capability of processing complex and continuous varying visual stimuli. The ability of motion detection must be undoubtedly mentioned among the most powerful faculties of Natural Visual Systems. Trying to emulate their astonishing performances represents a real challenge.

Further, this interest is also prompted by the increasing number of potential applications, such as smart video safety and video surveillance, intelligent gestural user-computer interfaces, orthopedic therapy and athlete training, or automatic content annotation.

Any proposal must deal with multiple targets whose dynamics are unknown and highly non-linear. Scene conditions are uncontrolled, unknown, and evolve over time due to changing illumination and weather, or moved objects. Targets move through scenes with complex clutter, which may mimic their appearances. Their trajectories are expected to intersect, thereby causing partial and complete occlusions. Finally, the approach should cope with heavy appearance and shape changes caused by the deformable and articulate nature of the targets, and variable illumination.

The goal is to implement and experimentally verify a novel image-based tracking architecture. It should be able to simultaneously perform a reliable tracking of multiple targets in unconstrained and dynamic open-world scenarios.

The architecture itself is considered as the main contribution: it introduces the necessary synergies to tackle such a inherently complex problem. Further, the different modules have been developed and improved —such segmentation [3], low-level tracking [2], high-level tracking [5], and event management [4].

2 A Framework to Human Sequence Evaluation

Human-Sequence Evaluation (HSE) defines a complete Cognitive Vision System which transforms image values into semantic descriptions of human behaviour by performing multiple bottom-up and top-down processes [1]. Its aim goes far beyond detecting, tracking and identifying the actions being performed: its goal is to apply cognition methodologies to understand human behaviour in image sequences.

The implementation of HSE involves three cooperating tasks: (i) the obtention of a dynamic description of the observed human motion; (ii) the transformation of these quantitative parameters into logic predicates; and (iii) the communication of the

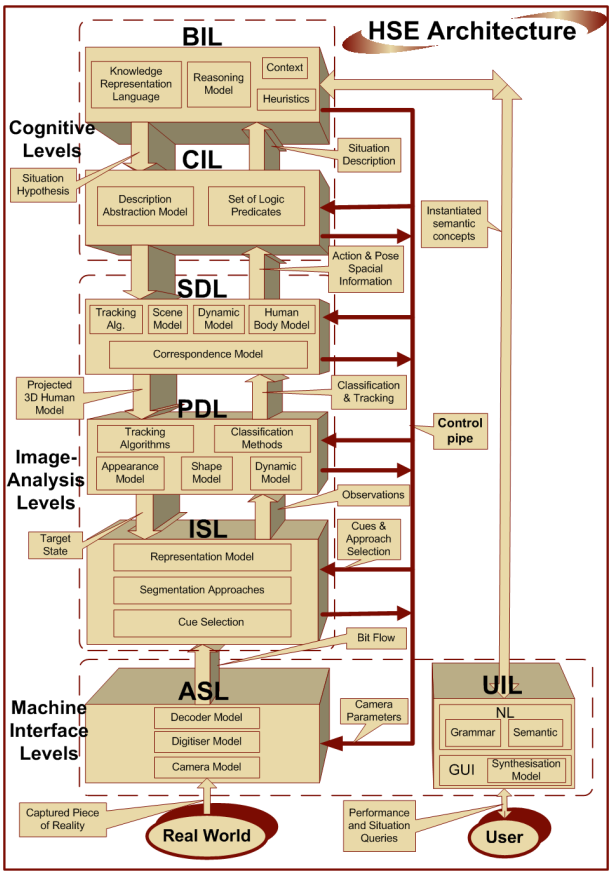


Figure 1: HSE framework evolved from [1]. Each level performs some general task such as providing a machine interface —ASL, UIL— processing and analysing the image sequence —ISL, PDL, SDL— and describing and reasoning over the obtained quantitative results —CIL, BIL.

obtained results to an human user. Multiple issues are then demanded: (i) active video camera control, (ii) target segmentation, (iii) robust and accurate MTT, (iv) target classification, (v) posture and action recognition, (vi) facial expression analysis, (vii) behaviour understanding, and (viii) communication to operators.

Due to this complexity, an HSE system is here presented as a structured framework, see Fig. 1. Levels are defined according to main functionalities. The whole structure is highly interconnected, and each level receives inputs from higher and lower ones, providing the system with redundancy. The inter-level communication can be seen in three different ways: (i) a data stream is provided to the higher levels by lower ones including all the results obtained in the bottom-up process; (ii) higher levels feed back the lower ones in a top-down process, so that the whole procedure can be enhanced; and (iii) higher levels can act on the lower ones by tuning the parameters, and selecting different operation modes, models or approaches depending on what is known about the current scene, and what goals are pursued.

The rest of this work is focused on the ISL, PDL and CIL within the HSE framework. Detection, estimation and adaptation tasks are here addressed.

3 System Architecture

Non-supervised MTT involves such a complexity that leads to propose a principled architecture. Reliable target segmentation is critical in every tracking system to achieve an accurate feature extraction without considering any prior knowledge about potential targets. However, complex interacting agents who move through cluttered environments require high-level analysis.

Our proposal combines both bottom-up and top-down approaches in a modular and hierarchically-organised architecture, see Fig. 2. A set of co-operating modules are distributed through three levels. Each level is defined according to the different tasks to be performed: Target Detection, LLT, and HLT.

The different modules take part in both bottom-up and top-down processes. The former provides the system with capabilities for initialisation, error-recovering and simultaneous modelling and tracking. The latter allows using detailed models according to a high-level event interpretation to perform tracking following a MBT or ABT approach.

These concurrent processes are allowed due to the fact that in the proposed architecture the tracking task is split into two levels: the lower one, which is based on short-term blob trackers, and a higher one, based on long-term target trackers. The latter has a crucial importance: it automatically builds and tunes multiple appearance colour models, manages the events in which the target is involved, and selects the most appropriate tracking approach according to these.

A complex event management is performed. Multiple-target interaction events, and a proper scheme for tracker instantiation and removal according to scene events, are considered. This allows the system to switch between both operation modes, and tackle open-world applications.

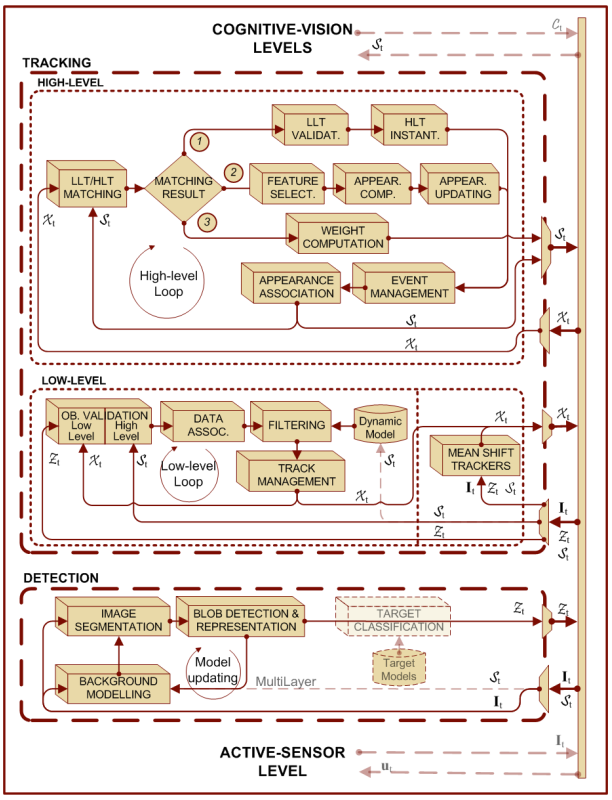


Figure 2: Tracking architecture. $\mathbf{I}_t$ represents the current frame; the observation, LLT and HLT data structures are denoted by $\mathcal{Z}_t$, $\mathcal{X}_t$ and $\mathcal{S}_t$ respectively; $\mathbf{u}_t$ represents a vector of potential system control signals, while $\mathcal{C}_t$ refers to high-level information. Ongoing and future-planned modules are shown in transparent dash lines.

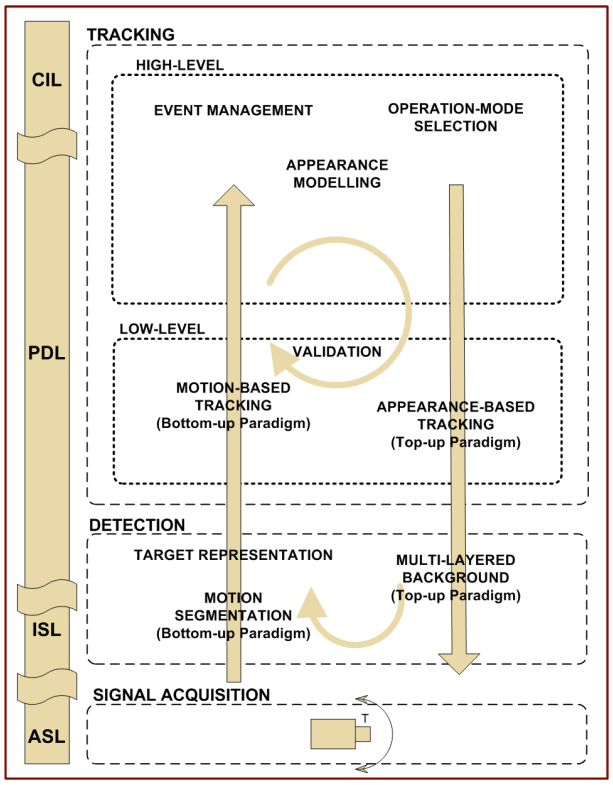


Figure 3: Relations between the HSE framework and the proposed tracking architecture. See text for details.

The approach copes with clutter distracters by selecting the most convenient colour-related features. A set of appearance models is continuously conformed, smoothed and updated. Thus, multiple targets are represented using several models for each of them, while they are simultaneously being tracked. Further, colour information relative to the target background and other close targets is used to tune the appearance models.

This tracking architecture is a part of the complex HSE framework, see in Fig 3. Segmentation tasks within the Detection Level correspond to ISL; target detection and classification, as well as LLT, and appearance representation within the HLT belong to PDL; and event management, operation-mode selection, and other HLT tasks are assimilated to CIL. The global position, shape and appearance of all targets within the scene is fed forward by this system.

4 Natural Paradigm Foundation

In a natural paradigm, visual-stimuli processing is divided into two categories: on the one hand, bottom-up or pre-attentive processes carry out raw data processing without high-level, a-priori learnt information; on the other hand, top-down or attentive processes perform goal-oriented tasks by making use of context and domain knowledge. Nevertheless, these two kind of processes are strongly linked, and they occur simultaneously in a closed loop. Further, the pre-attentive stage of vision performs the processing for different visual cues, such as motion or colour. This is done in a parallel and independent way. These results are fused in the attentive stage.

Hence, our proposed architecture follows this natural paradigm. The pre-attentive stage given by the LLT's provides a coarse localisation, while the attentive one performs an accurate tracking of those objects of interest.

(a) (b) (c)

Figure 4: Sample tracking results on *HERMES_Outdoor_Cam1* sequence. The dissolution of a non-detected group is correctly detected in (b); targets are successfully tracked through groups and occlusions; left objects are detected in (b), an correctly tracked after being picked up in (c).

LLT's are created at a initial level of abstraction. This step provides several advantages: (i) segmentation errors due to noise, camouflage, or the inclusion of shadows and reflections are reduced, thereby limiting potential spurious structural changes; (ii) the LLT's target representation can be handled by HLT's, thereby reducing the sensory gap between images and high-level abstractions; and (iii) the computational complexity is cut down by using a compact representation, which also removes confusing elements.

Thus, LLT's perform a rough tracking where detailed models are avoided. No appearance information is used, and events are not analysed. Subsequently, the system performs selective examinations of the tracked objects that draw its attention. Hence, HLT's build accurate appearance colour-based models, and analyse the events in which they take part in. This information is then used to act on the lower trackers, thereby yielding a closed-loop system.

The operation modes follow also the paradigm of first-order and second-order motion perception. While the former is performed by detecting changes in a particular point of the retina, the latter depends on moving blobs defined in terms of contrast or texture.

5 Experimental Results

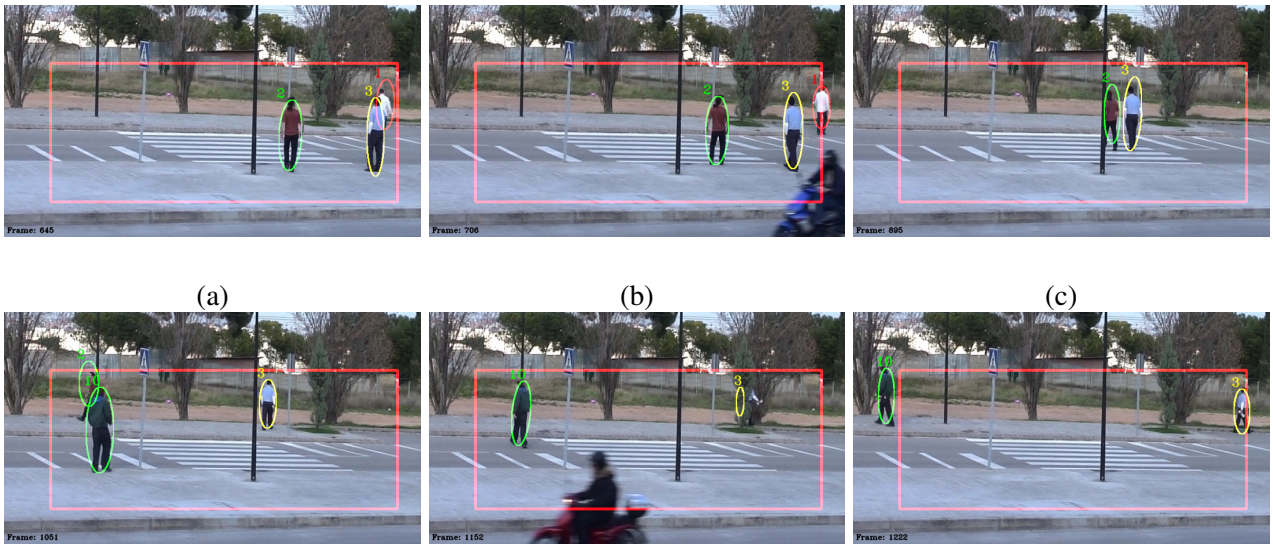
The performance of our system has been tested using sequences taken from both public well-known databases, and own ones. Successful tracking results have been achieved in all processed sequences[1], see Figs. 4, 5, 6.

Further, a ground-truth annotation tool has been developed, and the interaction between human and computer is aided by using a pen tablet. Thus, foreground regions can be annotated, visualised and edited. Targets are labelled, and visible and occluded regions are pointed out, as well as the head and feet.

Ground-truth events are confronted with computed ones, see Table 1. Events are correctly detected, albeit hardly ever occur at the exactly same time instant. This issue is of course sensitive to location estimation errors of a few pixels. However, target1 does not keep its ID after leaving the bag, due to major shape and appearance changes, and two new trackers are instantiated. Hence, target1 is referred as target 4 after bag is left. Consequently, subsequent tracker instantiations have the labels shifted.

Further, several trajectory indicators over the tracked targets are computed and presented in Table 2. Every time a new blob is detected, a LLT is instantiated. This usually happens when targets merge into groups, they dissolve themselves, or targets undergo significant changes due to camouflage, occlusions, etc. Thus, the number of LLT's is much higher than the number of targets in every analysed sequence. When a LLT become stable, a HLT is created and associated with it. These are hopefully subsequently associated with the HLT that is already tracking the target. In this case, the target identity is not broken. When this process last more than one frame, the identity is temporarily broken. Since a

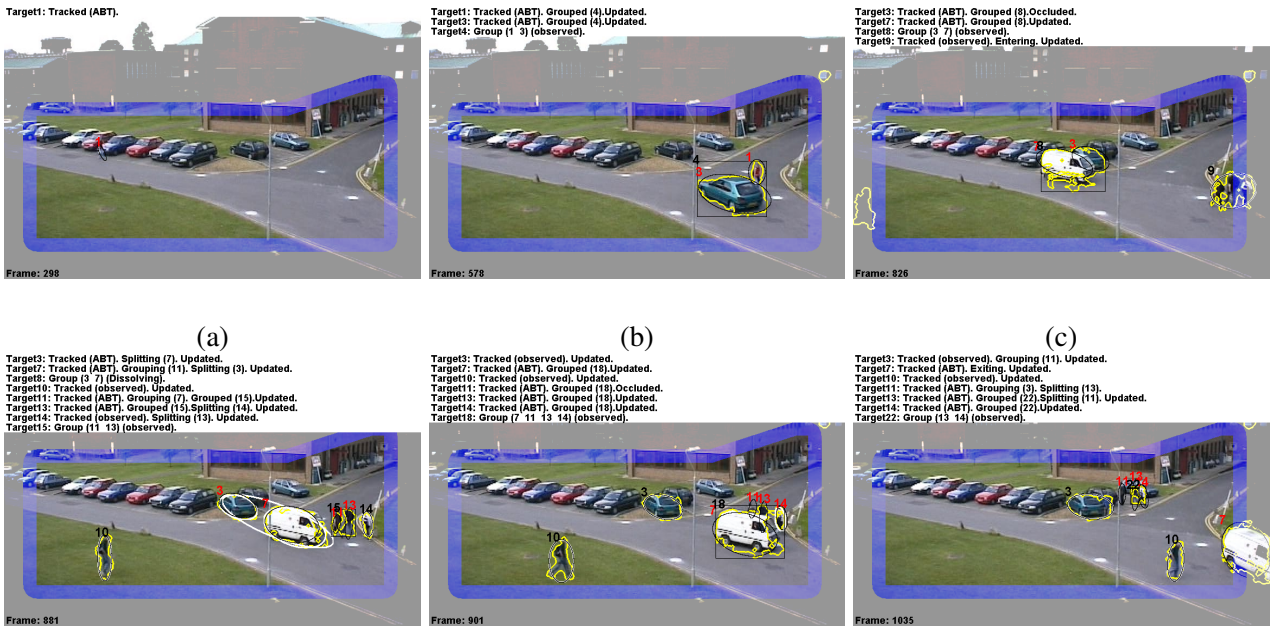
[1]The reader is encouraged to see the sequences at `http://iselab.cvc.uab.es/?q=agent_motion`



(a) (b) (c)

(d) (e) (f)

Figure 5: Sample tracking results on *CVC_Zebra1* sequence. Targets are successfully tracked despite mutual occlusions in (a) and (d), or occlusions with the background in (c) and (e); interaction and scene events are correctly inferred.



(a) (b) (c)

(d) (e) (f)

Figure 6: Sample tracking results on the *PETS_DATASET1_TESTING_CAMERA1* sequence. Targets are tracked despite no segmentation is available in (a), a single blob is obtained for the group in (b), (d), or they are heavily occluded in (e); multiple simultaneous events are correctly inferred, such as target 13 is grouped in group 15 while splitting from 14 in (d).

HLT is created after the event is over, together with the fact that HLT are also instantiated to track groups, the number of HLT's is higher than the actual number of targets, even if the identities are correctly kept.

Table 1: Annotated and computed events.

Event (t, ID)	Computed event
observed (550, 1)	observed (550, 1)
entering (629, 2)	entering (629, 2)
—	group dissolv. (655, 1)
split. from 3 (662, 1)	split. from 3 (655, 4)
split. from 1 (662, 3)	split. from 4 (655, 3)
group. with 2 (681, 1)	group. with 2 (682, 4)
group. with 1 (681, 2)	group. with 4 (682, 2)
grouped in 4 (689, 1)	grouped in 5 (697, 4)
grouped in 4 (689, 3)	grouped in 5 (697, 3)
group: 1 & 2 (689, 5)	group: 4 & 2 (697, 5)

Table 2: Trajectory Measures.

	CAVIAR	PETS	HERMES
Targets	2	8	8
LLT	8	78	86
HLT (tgs)	4	28	36
HLT (grs)	1	13	11
Broken ID	0	0	1
FP	0	0	2
FN	0	0	0

The permanent broken ID, and the false positives are due to *ghosts* yielded by a non-detected motionless car which starts motion.

6 Concluding Remarks

In this work a principled and structured system is presented in an attempt to take a step towards solving the numerous difficulties which appear in unconstrained tracking applications. The system here proposed implements a hierarchical but collaborative architecture, in which each level is composed of several modules which are devoted to specific tasks. Therefore, albeit the different modules have been here developed or improved, we consider the architecture itself as the main contribution: it introduces the synergies between the algorithms which permit to tackle a problem with such an inherent complexity.

This structured framework combines in a principled way both bottom-up and top-down tracking approaches: each level feeds the higher one with its computed results, and is itself fed back with high-level results. In this way, by taking advantage of both approaches, the system is allowed to benefit from bottom-up capabilities, such as simultaneous modelling and tracking without making used of a-priori knowledge; but also, high-level analysis is performed, granting accurately tuned models, and proper operation-mode selection. In addition, each level has an internal loop which also provides the system with adaptive capabilities by updating the background module, making use of the knowledge about existing tracks, or selecting the most appropriate approach according to the events in which the targets are involved.

Future research will be focused on enhancing target representation by including structure components —such as body-part histograms— and shape cues — such as SIFT descriptors. Further, the system will benefit from high-level information provided by the cognitive levels.

Acknowledgements. This work was supported by the Catalan Research Agency (AGAUR), the EC grants IST-027110 for the HERMES project and IST-045547 for the VIDI-Video project, and the Spanish MEC under projects TIN2006-14606 and DPI-2004-5414. Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

[1] J. Gonzàlez. *Human Sequence Evaluation: The Key-frame Approach*. PhD thesis, UAB, Spain, 2004.

[2] J. Gonzàlez, D. Rowe, J. Andrade, and J.J. Villanueva. Efficient Management of Multiple Agent Tracking through Observation Handling. In *6th VIIP*, pages 585–590, 2006.

[3] I. Huerta, D. Rowe, M. Mozerov, and J. Gonzàlez. Improving Background Subtraction based on a Casuistry of Colour-Motion Segmentation Problems. In *3rd IbPRIA*, volume 2, pages 475–482, 2007.

[4] D. Rowe, J. Gonzàlez, I. Huerta, and J.J. Villanueva. On Reasoning over Tracking Events. In *15th SCIA*, pages 502–511, 2007.

[5] D. Rowe, I. Reid, J. Gonzàlez, and J.J. Villanueva. Unconstrained Multiple-people Tracking. In *28th DAGM*, pages 505–514, 2006.