The SmarTrack Project at FBK: Past and Ongoing Efforts on People Tracking for Surveillance and Monitoring

Oswald Lanz and Stefano Messelodi

Abstract Progress in computer vision research is reshaping the video surveillance sector: high-tech companies are starting to offer CCTV based systems now empowered with Video Analytics, i.e. with software solutions able to generate meaningful alerts by analyzing video feeds. Most surveillance applications target people as their subject of study, where they move, how they behave, and what they carry (or leave unattended); people tracking is therefore becoming a ever more important functionality for new generation technology. In this document we summarize our efforts in realizing SmarTrack, a multi camera people tracker developed by FBK over the last few years. We detail main research results, development efforts and applications, and present current and future research directions.

1 Introduction

Ambient Intelligence is a paradigm that refers to technology-equipped environments that are sensitive and responsive to the presence and actions of people. What to *sense* about the environment, and how to *react* on it depends on the specific application domain: Security & Surveillance, Ambient Assisted Living, Retail Monitoring, Sports Analysis, etc., have different requirements in terms of system feedback; however, they all share the need to gather information about peoples' locations over time, employing non-intrusive sensors. The SmarTrack project at FBK addresses this need, and is an effort to realize a software component to track people using a set of cameras.

Oswald Lanz
FBK, Via Sommarive 18, Povo di Trento (TN), ITALY, e-mail: lanz@fbk.eu

Stefano Messelodi
FBK, Via Sommarive 18, Povo di Trento (TN), ITALY, e-mail: messelod@fbk.eu

Object tracking is a well studied topic in Computer Vision (see [2] for a survey) and proposed solutions differ largely in type of environment they are designed to operate on (indoor vs. outdoor, single camera vs. multi-camera) and adopted metodology (model-based vs. data driven, distributed vs. centralized). SmarTrack targets persistent people tracking in structured environments employing a distributed sensing and processing infrastructure. A prototype scenario is that of smart museums: to operate reliably, such system should be

- robust to persistent occlusions. People often move in groups, and some of them may be occluded in one or more camera views for a significant amount of time (e.g. during a museum visit). Monitoring a subject across such events relies either on re-identification, or on explicit occlusion handling (we investigated this latter).
- scalable. Solutions should operate in real time on a resource constrained infrastructure and scale to environments with complex topology. In particular, our focus is on limitations of traditional tracking characterized by passive sensing and limited adaptation.
- flexible. Ideally, solutions should be customisable to (jointly) track a selected set of objects of various kinds (people, cars, luggage) and operate on different modalities (including 3D vision, infrared, but also acoustic sensing).

With these goals in mind, in the following we present the SmarTrack approach with main research results at its current state (Sec. 2), describe applications and related projects (Sec. 3), and give an outlook on current and ongoing efforts (Sec. 4).

2 SmarTrack: Approach and Main Results

In SmarTrack we adopt a probabilistic Bayesian framework and represent both target motion and appearance in 3D space. This allows to embed explicit models of target behaviour and sensor mapping (see Fig. 1 for the likelihood model used by SmarTrack) based on physics principles, including geometric constraints and visual occlusion, and to account for ambiguity and uncertainty which is often inherently present in the data. Since the targets are represented in 3D space, motion, appearance and sensor mapping models can be designed to track various objects, using various features (color, texture, edges, flow), at various resolutions (position, pose, articulated motion), in various modalities (passive vision, 3D vision, audio). Examples are given in Sec.3, while the contributions detailed in this section abstract from these choices.

While Bayesian approaches are typically more demanding in terms of computations than their data-driven counterparts, we show that in the case of visual tracking with explicit occlusion handling such approach can be effectively scaled to the number of targets (Sec. 2.1), and that the same underlying model can be adopted to solve the detection problem, via likelihood inversion (Sec. 2.2). Further savings can be achieved by continuously adapting the resolution of probabilistic estimates to their uncertainty, a result that allows to consistently manage the trade-off between

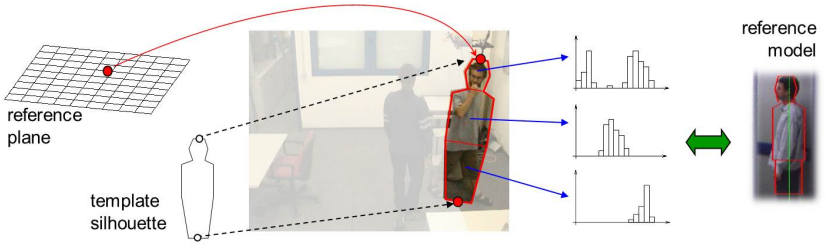

Fig. 1 Likelihood model of SmarTrack. Given a hypothesis (i.e. a particle), a 3D shape model is projected onto the image and color histograms are extracted from the projection. To assign a score, these histograms are then compared to a previously acquired model of the target (Sec. 2.2).

robustness and efficiency in a particle filter implementation (Sec. 2.3). To highlight the flexibility of the approach, we show in Sec. 3 how pose estimation and how multi modal (audio-video) information can be integrated in an easy and robust way.

2.1 Tracking multiple occluding bodies: the HJS approach

Independent tracking of individual bodies fails in the presence of occlusions, where the disappearance of a target cannot be explained if not in relationship with the other targets. On the other hand, describing the dynamics of the different bodies with a joint model requires a representation size and computational cost that grow exponentially with the number of bodies. To allow for a tractable solution, the Hybrid Joint-Separable (HJS) model has been proposed [1]. Instead of maintaining distributions in their joint form, a factorial representation in form of product of marginals is estimated recursively.

As our main contribution in this field, in [3] this model is adopted to solve the multi-target tracking problem with a occlusion robust extension of the likelihood in Fig. 1. The salient property of this approach is that it allows for tractable inference which is understood and theoretically grounded, and that it scales to input complexity, i.e. number of targets, K: its computational complexity has a quadratic upper bound $O(N \cdot K^2)$, compared to $O(N^K)$ of traditional algorithms (N is the number of particles per target, e.g. $N = 200$). The approach allows also to embed a MRF motion model to account for pairwise interaction [3]: with the MRF implementing spatial exclusion (i.e. a constraint that two targets can not share the same location) the tracker is able to resolve severe occlusions among multiple targets with similar appearance (avoiding coalescence). Fig. 2 shows some frames of a difficult sequence successfully tracked with the extension proposed in [6]. The method is patent pending.

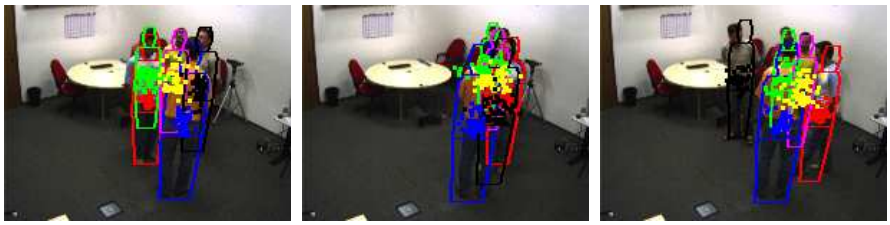

Fig. 2 Five targets successully tracked using one camera, under severe occlusions, in a lab setting.

2.2 Target detection via likelihood inversion

The effectiveness of such approach depends on the appearance model upon which the target is re-localized from frame to frame and, importantly, how well such a model is initialized (and updated over time; this is still an open issue for us, see Sec. 4). Ideally, automatic initialization should be triggered by a detection process that searches for new objects *through the view of the re-localization process* for which the appearance model has to be calibrated. In SmarTrack we have derived a detection method that builds upon the same likelihood formulation, but, instead of a color signature used for tracking, relies on motion edges to render its features target independent.

Apparent motion can be quickly extracted using image differencing and resembles the silhouette contour of a moving object with good approximation. Therefore, the shape projection method designed for tracking (Fig. 1) can be used for detection in a search based fashion via contour matching (i.e by measuring how well the shape projection outline is covered by motion edges). Note that the search here must be performed over the joint space of 3D location and shape dimensions. To get a scalable solution [7] we

- [off-line] compute, for each pixel, a representation of the set of object configurations whose projected shape outline maps onto that pixel (done in a calibration step, before tracking), and
- [on-line] assemble, for each tracker iteration, a detection probability map by fusing the set of representations indexed by motion edges that are not explained by other tracked targets; a prominent peak in the map triggers the initialization of a new track.

The method can account for occlusions generated by tracked targets: it has been derived by inversion of the HJS likelihood [7]. Fig. 3 shows the successful detection of a person in a occluded context. The method is patent pending.

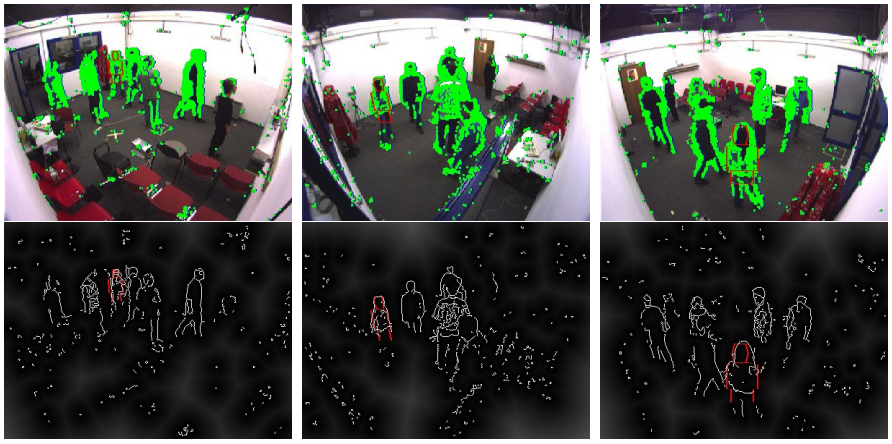

Fig. 3 Detection in a challenging situation (6 out of the 7 people were already tracked).

2.3 Adapting the number of hypotheses to propagate

SmarTrack implements the HJS method via particle filtering, i.e by propagating a compressed representation of the posteriors by means of particles, a set of states with high probability. The number N of particles to propagate is a parameter that regulates the trade-off between robustness (larger N means denser representation of the posteriors) and efficiency (low N means less likelihoods to compute). The uncertainty of an estimation process depends on many factors: the impact of clutter and occlusions on it changes with the position of the targets. To track robustly with minimal resources the number of particles N should therefore be adapted over time.

The work in [5] presents a method for adapting N on-line, during tracking, to allocate more particles when uncertainty is high while saving resources otherwise. The key idea is to select the number of samples necessary to populate the high probability regions with a predefined density (a parameter of the method). In information theory the notion of high probability region is formalized through the typical set, and the Asymptotic Equipartition Property (AEP) theorem states that the volume of such a set is related directly to the entropy of the pdf it originates from. Based on this property, a scheme for adapting the number N of hypotheses propagated by SmarTrack has been derived: at each tracker iteration the scheme computes the amount of uncertainty carried by the particles by means of the AEP and entropy estimation (via a kernel density approach), and adapts N accordingly.

2.4 SmarTrack at IST 2006

An early version of SmarTrack has been showcased at the IST event in 2006, earning an award. During the event more than 900 people visiting the hosting CHIL

Fig. 4 The CHIL Project booth at IST'06 and SmarTrack's view of it during the event.

Project stand have been tracked (Fig. 4). By using SmarTrack as an interface, visitors were able interact with a tabletop device located in the booth, displaying a shared workspace. An automatically generated personalized report has been handed out to each visitor containing a plot of his path through the booth and a description of the demo he was most interested in (i.e. where, according to collected tracks, he has spend most of the time).

3 Applications and Related Projects

3.1 Monitoring the elderly at home and during care

With the ever aging population there is a real need for technology that can be integrated cost effectively into elderly's homes in order to make them feel in care while more comfortable about remaining in this familiar environment. Also, there is an interest in developing systems to assist caregivers in a nursing environment. In both cases such systems should support the elderly during its everyday activities in their living environment, monitor their behaviour to measure well-being, and be able to detect situations of potential danger.

In NETCARITY (an EU FP6 project, `http://www.netcarity.org/`) and ACUBE (a project sponsored by the Autonomous Province of Trento PAT, `http://acube.fbk.eu/`) we use SmarTrack to provide tracking capabilities to such a system. A significant effort is devoted to the development of a prototype to be integrated in a multi-sensorial system with high level functionalities targeting end user needs, and to deploy it in a real environment, with real users. In ACUBE three deployments are foreseen: in a domotic lab (for demo purposes), a domotic house (for integration tests) and in a pilot site (with Alzheimer patients).

3.2 Tracking position and pose to monitor attention

For monitoring purposes it is often desirable to collect additional information on the subject of interest, such as where a person is looking at, e.g. to analyze customer behaviour in a shopping mall. Such information can be acquired from video recordings by means of tracking the spatial position and head orientation of people. Its visual attention can then be logged, by intersecting the viewing cone of the subject (rooted at the head position and oriented according to the estimated pose) with the trajectories of other people and the location of objects collected in its surrounding environment.

Within CHIL (an EU FP6 project, `http://chil.server.de/`) we have build on SmarTrack to extract such information, including head pose, from far-field recordings in a Smart Room environment. The goal, among others, was to provide cues in support to automated analysis of meetings (e.g. tools to determine the leader of a discussion, who is interacting, paying attention, and who not, etc.). To obtain such information we expanded the estimation space of SmarTrack to include the horizontal orientation of a target and re-implemented the likelihood function (Fig. 1) to get a sensor mapping method that is sensitive to the additional dimension [4]. With this customization the tracker estimates position and head pose jointly from low resolution images, where pose estimation techniques based facial features cannot be applied. Fig. 5 shows the output of SmarTrack on some frames of a sequence used to produce an automatic transcription of the interaction of a child with an attention object, the ball, which was also simultaneously tracked in the 3D space [8]. Note that also body orientation and inclination of the study subject was estimated (again by expanding the estimation space and the shape projection method, as for the ball).

Within PUMALAB (an FBK internal project) we also applied SmarTrack to investigate on the relationship between proxemics, visual attention and personality traits during interaction. In [9] we showed that from data recorded during natural interaction (a party) a psychologically grounded model of behaviour emerged in the tracker's output, and that such model can be calibrated automatically to each individual to capture inter-personal variations that correlate with personality traits.

3.3 Integrating audio to track speech activity

Automatic analysis of interactive people behaviour is an emerging field where significant efforts of the audio and image processing communities converge. Reports about a persons visual attention and speaking activity may be used to study its behaviour during meetings, but also during social interactions, accounting for its various modalities.

Within CHIL and PUMALAB we have investigated how information from multiple and heterogeneous sensors can be integrated in SmarTrack, on the example of audio visual tracking. This is done by expanding the estimation space with a speech activity flag and implementing an sensor mapping function to sense it [10]: given

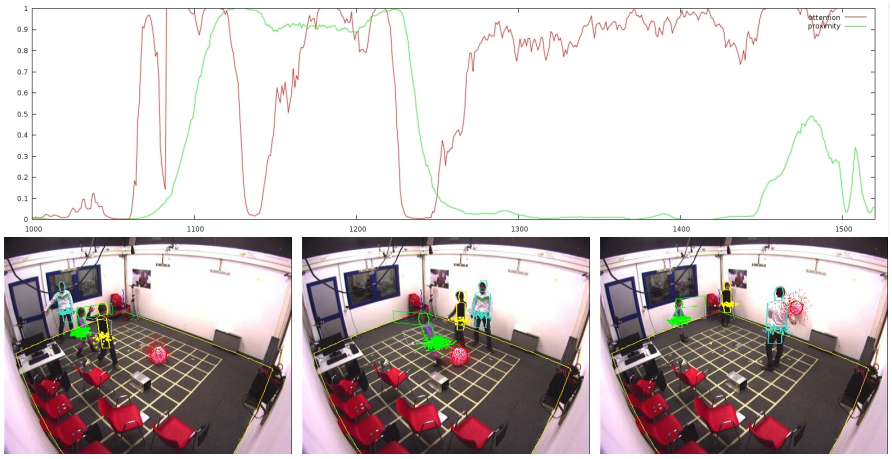

Fig. 5 Real time transcription of visual attention: visual focus (in red) and proximity (green) of a study subject (green) towards the attention object (purple) over time, and raw output of SmarTrack on example frames from which such features were extracted.

a particle with active flag we (i) compute the theoretical time delay with which an acoustic signal emitted at the particles position hypothesis would arrive in two microphones placed at a known distance (we assume that the distributed microphone array is geometrically calibrated) and (ii) verify how well and with which energy the captured signals correlate under that time shift. By integrating such measure as an additional likelihood SmarTrack is able to detect when and in which direction a tracked person is speaking and, importantly, who among the tracked ones is speaking. We want to use this additional cue to progress on the automatic analysis of natural interaction (previous section). Also, we are currently investigating how to integrate 3D information provided by the Kinect sensor in a similar way, by extending the image sensor mapping function.

4 Current and Ongoing Research

4.1 Tracking in unevenly illuminated scenes

From installations in the various pilot sites (of ACUBE, in particular) it has emerged that, in order to operate reliably in unevenly illuminated scenes (e.g. in a windowed home environment), the color model used to re-localize a target over time must be adapted during tracking to match the illumination conditions at its current location.

Until now, our activities to this regard have focused on the semi-automatic acquisition of illumination maps encoding local illumination conditions in a off-line calibration step, and to use them in SmarTrack to remap, at each iteration, the ap-

pearance models of each target into a locally normalized color space [11]. When the illumination conditions chance over time these maps have to be updated: our goal is to extend the HJS model to learn and update non-parametric representations of such maps using people as 'sensors' to collect evidence for adaptation, possibly in a multi-modal setting (including acoustic sensing and 3D information provided by the Kinect sensor). Such method might also be applied to learn a model of the acoustic distortion function (essential in a reverberant environment with complex absorbation pattern) for improved multi-modal speaker localization, and, more generally, to extract persistent patterns that can be modeled in the tracking space and sensed in the various modalities (e.g. [12]).

4.2 Boosting scalability through active sensing

Until recently, object tracking has mostly been considered as a pure estimation problem, i.e. as the problem of inferring information about the targets' location (and/or movement, orientation, pose, etc.) over time from *given sensor data*, i.e. from observations acquired by an independent process. A notable exception comes from robotics, where the possibility of purposely driving a sensor-equipped robotic platform (thus controlling the sensing process) for better inference to solve the simultaneous localization and mapping (SLAM) problem has got large attention earlier. However, there the focus was on the spatial dimension of active sensing, i.e. where to (move to) take the next observation in compliance with the constraints of the resource, i.e. with the controllable but limited mobility of the platform. The temporal and multi-modal dimensions, i.e. when to take the next observation for inference, and from which of a set of available and possibly heterogeneous sensors, become of evident importance if additional limitations are to be considered, possibly of critical importance if scalability is an issue.

In this context, our investigations are twofold [13]; we aim at developing:

- Task-driven polling strategies that control the sensing process in order to minimize the number of observations to be (transmitted and) processed while maximizing their expected impact on the estimation process;
- Parameter adaptation techniques for the estimation process (e.g. number of propagated hypotheses in a particle filter tracking framework, progressing on [5]) that re-allocate computational resources dynamically and opportunistically, according to task complexity and measured evidence.

The first objective has obvious implications in terms of energy consumption at sensor side (the sensor is queried only if this is requested by the estimation process), and in terms of network resources if data has to be transferred to a processing node (e.g. in a centralized architecture). The second objective impacts on the amount of processing to be performed on each requested observation, which may even be carried out locally on the sensor, impacting on energy consumption as well. Effective solutions to both problems are thus expected to have a significant impact on a mon-

itoring system's throughput and scalability, potentially enabling them to operate on previously uncovered scenarios.

Acknowledgements Activity partially funded under projects: PEACH (Grand Project, Autonomous Province of Trento PAT), CHIL (EU FP6), NETCARITY (EU FP6), ACUBE (Grand Project, PAT), PUMALAB (FBK internal project, PAT).

References

1. Yilmaz, A., Javed, O., and Shah, M. (2006): Object tracking: A survey. ACM Comput. Surv. 38, 4, Article 13, 2006.
2. Lanz, O. and Manduchi, R. (2005): Hybrid Joint-Separable Multibody Tracking. In IEEE CVPR, 2005.
3. Lanz, O. (2006): Approximate Bayesian Multibody Tracking. IEEE Trans. PAMI Vol. 28, No. 9, 2006.
4. Lanz, O. and Brunelli, R. (2006): Dynamic Head Location and Pose from Video. In IEEE MFI, 2006.
5. Lanz, O. (2007): An Information Theoretic Rule for Sample Size Adaptation in Particle Filtering. In ICIAP, 2007.
6. Lanz, O. (2009): A HJS Filter to Track Visually Interacting Targets. In IEEE ICASSP, 2009.
7. Lanz, O. and Messelodi, S. (2009): A Sampling Algorithm for Occlusion Robust Multi Target Detection. In IEEE AVSS, 2009.
8. Poesio, M., Baroni, M., Lanz, O., Lenci, A., Potamianos, A., Schuetze, H., Schulte im Walde, S. and Surian, L. (2010): BabyExp: Constructing a huge multimodal resource to acquire commonsense knowledge like children do. In LREC, 2010.
9. Zen, G., Lepri, O., Ricci, E. and Lanz, O. (2010): Space Speaks - Towards Socially and Personality Aware Visual Surveillance. In ACM MPVA Workshop, 2010.
10. Brutti, A. and Lanz, O. (2010): A joint particle filter to track the position and head orientation of people using audio visual cues. In EUSIPCO, 2010.
11. Zen, G., Lanz, O., Messelodi, S. and Ricci, E. (2010): Tracking Multiple People with Illumination Maps. In ICPR, 2010.
12. Ricci, E., Tobia, F. and Zen, G. (2010): Learning Pedestrian Trajectories with Kernels. In ICPR, 2010.
13. Hu, T. and Lanz, O. (2011): Dynamic Resource Allocation for Probabilistic Tracking via Attentive Sensing and Sampling. In RAWSNETS Workshop, 2011.