Pose-Configurable Generic Tracking of Elongated Objects

Daniel Wesierski
Multimedia Systems Department
Gdansk University of Technology
daniel.wesierski@pg.gda.pl

Patrick Horain
Departement Electronique et Physique
Institut Mines-Telecom / Telecom SudParis
patrick.horain@telecom-sudaris.eu

Abstract

Elongated objects have various shapes and can shift, rotate, change scale, and be rigid or deform by flexing, articulating, and vibrating, with examples as varied as a glass bottle, a robotic arm, a surgical suture, a finger pair, a tram, and a guitar string. This generally makes tracking of poses of elongated objects very challenging.

We describe a unified, configurable framework for tracking the pose of elongated objects, which move in the image plane and extend over the image region. Our method strives for simplicity, versatility, and efficiency. The object is decomposed into a chained assembly of segments of multiple parts that are arranged under a hierarchy of tailored spatio-temporal constraints. In this hierarchy, segments can rescale independently while their elasticity is controlled with global orientations and local distances.

While the trend in tracking is to design complex, structure-free algorithms that update object appearance on-line, we show that our tracker, with the novel but remarkably simple, structured organization of parts with constant appearance, reaches or improves state-of-the-art performance. Most importantly, our model can be easily configured to track exact pose of arbitrary, elongated objects in the image plane. The tracker can run up to 100 `fps` *on a desktop PC, yet the computation time scales linearly with the number of object parts. To our knowledge, this is the first approach to generic tracking of elongated objects.*

1. Introduction

Elongated objects constitute a large, general class of structures that extend over image regions. They can move fast under varying illumination and occlusions, in clutter, and deform in the camera projective space due to relaxed rigidity or change in viewpoint. Yet, applications requiring pose tracking of elongated objects are various and span, *e.g.*, interactive video manipulation, telesurgery, gesture-based control, activity recognition, and animation. Hence, tracking elongated objects is a challenging but important task.

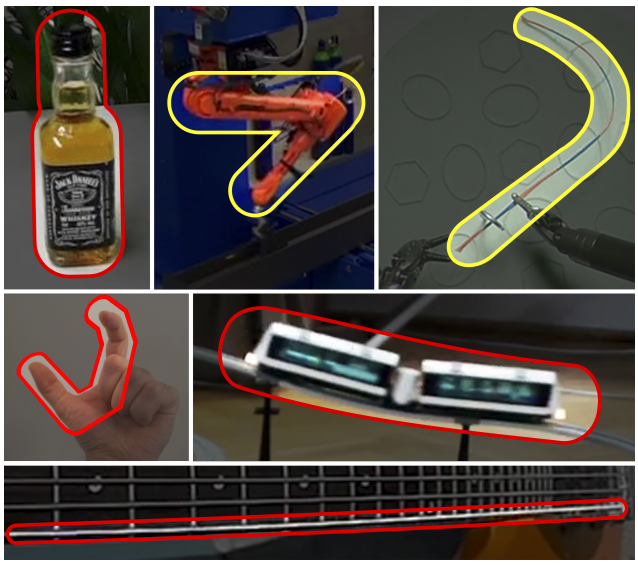

Figure 1. Our goal is to track with *one* algorithm poses of *plethora* of elongated objects varying in shape, motion, and rigidity. Our approach decomposes an elongated object into a chained assembly of segments of multiple parts that are arranged under a hierarchy of tailored spatio-temporal constraints leveraging local rigidity over object segments. As a result, we efficiently track elongated objects that can shift, rotate, change scale, and be rigid or deform by flexing, articulating, and vibrating.

However, an algorithm that tracks precisely, robustly, and rapidly a plethora of elongated objects varying in shape, motion, and rigidity has not been proposed thus far.

Dedicated trackers have made significant progress in specific, important areas (*e.g.*, surface deformations of human face [36], articulating tree-based human pose [28]). They can self-start but annotating training examples of all possible objects for learning spatially structured models is currently difficult. In contrast, structure-free, *generic* approaches, which are initialized simply by a single bounding-box, can localize arbitrary objects that are rigid [19, 34], deform less [5, 22, 33, 41], or more [6, 12, 23]. They build object appearance on-line but strive to be robust *against*

object deformations and thus neglect or filter out its pose. Arguably, the single bounding-box annotation scenario currently limits their applicability to elongated objects that occupy rather expanded image regions.

In view of this, the paper addresses a new problem of developing a generic system for pose-based tracking of elongated objects, which we conformably define as chain-like image structures. We position our approach between the structured and structure-free trackers by treating elongated objects as a structure of chained segments of parts with fixed appearance. Existing computer vision techniques, a pictorial structure, dynamic programming, and color histograms, are integrated into a new but simple model, which is composed of a hierarchy of spatio-temporal constraints with global orientations over the chained segments, contributing to model-based tracking. Notably, we introduce a generic, model-based tracker that admits a simple, one-shot configuration from annotated object parts in the first frame. Apart from its computational efficiency, it also tracks objects robustly against partial occlusions and local appearance changes due to spatial support through part-based structure and re-detects them after full occlusions due to temporal support through fixed appearance. Our system can be configured to efficiently estimate *detailed* pose trajectories of *elongated* objects, as *varied* as a rigid glass bottle, a flexing tram, a gesturing finger pair, an articulating robotic arm, a deformable surgical suture, and a vibrating guitar string, which extend over wider and very thin image regions, as depicted in Fig. 1.

We achieve this within a MAP-MRF setting of pictorial structures [10, 11] by developing a deformable model of chained parts that efficiently leverages object *local rigidity over spatio-temporal domain*. Specifically, the *fixed* appearance of each square-like part is represented by a color histogram, which has low computational cost, is invariant to scale change and to permutation of pixels. This means the pixels can evolve freely within object parts during tracking, so achieving robustness to rotation and to local deformations caused by moderate change in viewpoint. We then maintain spatial appearance of the whole object by decomposing it into a chained assembly of segments of multiple parts that are arranged under a hierarchy of tailored spatio-temporal constraints. We reference each segment of parts with an oriented polar coordinate system, effectively enforcing the spatial coherency of parts by promoting these part configurations that conform to the preferred relative angular deviations and distances over time.

Contributions: Our main contributions are: (i) a pose-configurable system for generic tracking of elongated objects, modeled with a hierarchy of tailored spatio-temporal constraints; (ii) demonstrating that a simple, structured organization of parts with fixed appearance leads to competitive performance with respect to structure-free, state-of-the-art methods that learn object appearance on-line. Our other contribution is to devise the new task of generic tracking of elongated objects having arbitrary shapes and motions. We also contribute by demonstrating that even though pictorial structures are usually considered slow [17], we integrate them into a hierarchical model that can register object pose up to speeds far exceeding real-time.

2. Related work

We review related work on region and part-based trackers of object poses, and other chain-based assemblies representing elongated image structures.

Region-based tracking: The seminal work of [8] proposed a mean-shift method that represented a non-rigid object by a color histogram, modulated with an ellipsoidal kernel. The tracker determined object location in real-time by mean-shifting the kernel in the gradient-ascending direction of the differentiated objective function. Owing to its simplicity, robustness, and speed, it has been popular and has evolved over the years [7, 14, 24]. In particular, [43] represents an elongated, rigid object by an asymmetric kernel and determines its location, scale, and orientation. However, these algorithms search locally (except [39]) requiring objects to move slowly. Also, they use a holistic appearance template that loses spatial information, reduces their robustness to occlusions [1], and renders them infeasible to track objects that deform heavily. Possibly, these types of objects may require a part-based approach [4].

Tracking by parts: Part-based trackers can represent objects locally. Thus, they can learn fewer background pixels that otherwise compromise the performance of holistic, bounding-box trackers [12]. However, they differ much in the mechanisms for assembling and matching the parts in the spatio-temporal domain [6]. For instance, parts described by fixed, gray histograms voted for object location in [1]. A human body was localized in [38] with several parts that were aggregated by greedy coverage of the foreground binary mask, obtained by graph-cuts. Kernels of parts were jointly mean-shifted in [9] to follow object deformations but required precomputing the subspace over their possible displacements on initial series of images to guide their joint convergence. Particle filtering [16] was used in [29] for probabilistic matching of several parts, defined by color histograms, which improved stability over the holistic template. The parts were linked rigidly, though, for efficient inference. Their unconstrained flexibility was then granted in [26] but through multi-stage, disjoint inference. Particle filters scale exponentially with the dimensionality of the search space, thus with the number of parts, and are approximative. On the other hand, the prominent pictorial structures [10, 11] have been used extensively in object tracking by approximating complete graphs with star graphs [3, 31] and with other tree graph extensions [42, 44].

The graphs are trained off-line for specific objects, but can explain heavy foreshortening [35] and scale linearly with object parts.

We aim at an efficient and precise framework to track elongated objects that can vary in number of parts by several orders of magnitude. In our setting, the primary advantage over particle filter and other pictorial structure trackers is that our tracker can render the global solution without approximative inference nor approximative object structure and its *joint* inference scales linearly with the number of parts. A chain-based pictorial structure thus appears natural to track elongated regions, and our approach generalizes to such structures of arbitrary rigidity in a computationally efficient manner.

Chain structures: Our proposal allows us to draw analogies to very influential snakes models of image contours [21], which can represent image structures with a chain graph [2, 15]. Snakes *actively* adapt to previously unseen contours to delineate object segments for shape registration. This is attractive, but they struggle on cluttered areas [37], so [32] uses region support. Essentially though, snakes use iterative matching methods with local search, thus additionally struggling with fast object displacements. Our work is also related in approach to [17, 40] that use a chained pictorial structure, and loosely related to [20] that iteratively infers on a dense graph by evaluating an ensemble of chains. However, [40] tracks non-deformable objects that shift and rotate, [20] requires a large set of training examples, and [17, 20] track object keypoints by filtering out object pose. Our approach generalizes to arbitrary, elongated objects (*e.g.*, curved) that shift, rotate, change scale, and undergo constrained or heavy deformations. It also learns object structure over region support with one-shot annotation, registers object pose explicitly, uses simple color histogram features to describe regions, and allows for finding the global solution in a single pass.

3. Approach

We develop a model-based approach that can track the motion of the pose of an arbitrary, elongated object in the image plane. We first partition an elongated object O_e into K segments $O_e = \{O_i\}_{i=1}^{K}$, as depicted in Fig. 2. Then, each segment i is partitioned further into k_i parts $O_i = \{p_{i,j}\}_{j=1}^{k_i}$ specified by square-like windows. Note that the parts need not be semantic nor have equal windows. We link the the parts with a chain graph $\mathcal{G}_c = (\mathcal{V}, \mathcal{E})$, where nodes $\mathcal{V}$ are associated with the parts and edges $\mathcal{E}$ are associated with the links between consecutive parts in the chain.

Each part $p_{i,j}$ is associated with an observed appearance feature $f_{i,j}$ and with hidden center location $l_{i,j}^t = [x_{i,j}^t\ ,\ y_{i,j}^t]^{\mathrm{T}}$ and scale $s_{i,j}^t$, forming random variable $p_{i,j}^t = [l_{i,j}^t\ ,\ s_{i,j}^t]^{\mathrm{T}}$. The variables are indexed with time $t = 1, \ldots$

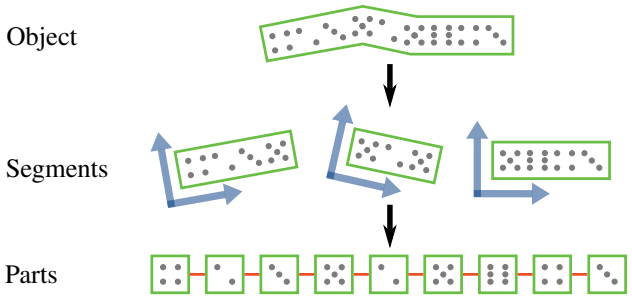

Figure 2. Model hierarchy, with an example of a deformable, elongated object, decomposed into K=3 segments that are referenced with planar coordinate systems. Each segment is split further into smaller parts, connected by pairwise distances to control its stretching and shrinking. Two segments share a part, which is anchored at their hinge, denoting heavy deformation (*e.g.*, articulation). The orientation of the coordinate system of each segment is estimated based on the tracked locations of the centers of the parts. The coordinate systems are used, in turn, to control the bending of each segment. In effect, the model captures deformations of the object and maintains its spatial coherence over time.

and we keep $f_{i,j}$ constant in this work. We index the initial frame with $t = 0$. The posterior over our pictorial structure of O_e^t at frame I^t yields:

$$P(O_e^t|I^t, O_e^{t-1}) \propto \tag{1}$$
$$\prod_{i=1}^{K}\prod_{j=1}^{k_i} \underbrace{P(I^t|p_{i,j}^t)}_{\text{Appearance term}} \prod_{j=1}^{k_i-1} \underbrace{P(p_{i,j}^t, p_{i,j+1}^t)}_{\text{Spatial term}} \underbrace{P(p_{i,j}^t, p_{i,j+1}^t|O_i^{t-1})}_{\text{Temporal term}},$$

where we set $p_{i,k_i}^t = p_{i+1,1}^t$, i.e. the last part of each segment is the first part of the next segment in the chain, so denoting a hinge. Thus, our graph has $|\mathcal{V}| = \sum_{i=1}^{K} k_i - (K-1)$ nodes resulting in $|\mathcal{V}||p|$-dimensional state space.

Fixed appearance: The appearance of each variable $p_{i,j}$ is simply captured with a normalized color histogram $f_{i,j} = h_{i,j}$. It takes the following form:

$$P(I^t|p_{i,j}^t) = \exp(-\frac{1}{\nu_{i,j}}\chi^2(h_{i,j}, h_{i,j,c}^t)\ ) \tag{2}$$

where $\chi^2(h_{i,j}, h_{i,j,c}^t)$ is the chi-square distance between the model histogram $h_{i,j}$, precomputed in the initial frame I^0, and the histogram $h_{i,j,c}^t$ at a candidate location and scale for part $p_{i,j}$ in the current frame I^t, with $\nu_{i,j}$ responsible for possible appearance variations. Note our approach is not limited to orientation invariant features though. As we update the orientation of segments during tracking, orientation variant features (*e.g.*, gradient orientations) could be updated accordingly [25].

The elongated segments O_i extend over rigid or elastic regions. Pictorial structures whether model whole segments

and search exhaustively for their orientations [10, 31], or split segments further into parts and model their constraints locally [42]. We also split segments into parts but model them hierarchically with spatio-temporal constraints, *i.e.* with local distances between parts and global orientations over segments to control their linear and angular deformations, respectively. Constraining each segment in a chain with global orientation allows to control its local rigidity without the need for higher order cliques in the graph, which is the key to fast inference. Borrowing terminology from automatic control, we consider the orientation to be a *slow-changing variable*, which, in turn, allows us to update it with one-frame lag without sacrificing the effectiveness of the approach. In this way, such a general, inertial temporal prior regularizes the dynamics of an object by favoring shift motion that is common during tracking [43].

Spatial prior: Neighboring parts in the i-th segment, $p^t_{i,j}$ and $p^t_{i,j+1}$, are constrained to lie within some predefined euclidean distance $d^t_{i,j}$ from each other, where:

$$d^t_{i,j} = \left\| l^t_{i,j} - l^t_{i,j+1} \right\|_2 \tag{3}$$

However, the changing scale of the object affects the distances, so we obtain:

$$P(p^t_{i,j}, p^t_{i,j+1}) = P(l^t_{i,j}, l^t_{i,j+1} | s^t_{i,j}, s^t_{i,j+1}) P(s^t_{i,j}, s^t_{i,j+1}) \tag{4}$$

For simplicity, we model the joint scale prior $P(s^t_{i,j}, s^t_{i,j+1})$ for each pair of parts in the chain as a uniform distribution. Hence, we omit it and reduce the spatial term only to $P(l^t_{i,j}, l^t_{i,j+1} | s^t_{i,j}, s^t_{i,j+1})$ as:

$$\begin{aligned} P(p^t_{i,j}, p^t_{i,j+1}) &\propto P(l^t_{i,j}, l^t_{i,j+1} | s^t_{i,j}, s^t_{i,j+1}) \\ &\propto \mathcal{N}(\, d^t_{i,j} \,;\, \rho^t_{i,j} \mu^{t-1}_{i,j;i,j+1}, (\rho^t_{i,j} \sigma^{t-1}_{i,j;i,j+1})^2 \,) \end{aligned} \tag{5}$$

The parameters $\mu^{t-1}_{i,j;i,j+1}$ and $\sigma^{t-1}_{i,j;i,j+1}$ in (5), computed in the previous frame, denote mean distance between locations $l^t_{i,j}$ and $l^t_{i,j+1}$ of two neighbor parts and its standard deviation, respectively. They are rescaled with $\rho^t_{i,j}$ to capture their dependence on the local scales of parts. Here, we simply set the rescaling factor as an arithmetic mean of these scales $\rho^t_{i,j} = \frac{1}{2}\,(s^t_{i,j} + s^t_{i,j+1})$.

Shift-geared temporal prior: We reference each segment i with a 2D coordinate system (CS), having initial orientation Θ^0_i w.r.t. the image coordinate system. This allows for determining local angular relations of the neighboring parts, which are defined by the angular offsets $\theta_{i,j;i,j+1}$ in the CS as `arccos` between the vector $[1, 0]$ (defined in the CS) and the normalized vector $l^0_{i,j+1} - l^0_{i,j}$. The bending of all the parts in the segment is then controlled during tracking with the temporal term as:

$$P(p^t_{i,j}, p^t_{i,j+1} | O^{t-1}_i) = \mathcal{M}(\theta^t_{i,j;i,j+1}; \theta_{i,j;i,j+1} + \Theta^{t-1}_i, \kappa_i) \tag{6}$$

where $\mathcal{M}$ denotes the von Mises distribution and κ_i denotes angular stiffness. The stiffness penalizes angular deviations from $\theta_{i,j;i,j+1}$ (with offset orientation Θ^{t-1}_i) caused by deformation and rotation of the segment. Therefore, our model favors such arrangements of parts of the segment, which maintain predefined geometrical configuration, presuming that the orientation Θ^{t-1}_i does not change much between successive frames. The stiffness parameters κ_i can be adjusted to account for the anticipated change in angular speed of Θ^{t-1}_i between frames.

Configuration: Our system admits a simple, intuitive procedure for configuring the pose of an elongated object O_e in the initial frame I^0. We: (1) split O_e into $|\mathcal{V}|$ parts $p^0_{i,j}$ by specifying their locations and sizes, (2) link neighbor parts with a chain $\mathcal{G}_c$, (3) and specify K segments of parts with their corresponding orientations Θ^0_i. Then, the features $f_{i,j}$, mean pairwise distances $\mu^0_{i,j;i,j+1}$, and angular offsets $\theta_{i,j;i,j+1}$ from Θ^0_i are computed from these one-shot annotations (as special case, straight objects enforce $\theta_{i,j;i,j+1} = 0$, independent of annotation).

Inference: We match our model (1) to each frame I^t by inferring on its negative log-posterior $-\log(P(O^t_e | I^t, O^{t-1}_e))$ with dynamic programming to obtain the MAP configuration of the elongated object $O^t_{e,MAP}$. The inference is fast and its complexity scales linearly with the number of object parts $|\mathcal{V}|$.

Update: The global scale s^t of the whole object is computed as the average over scales of all windows of parts and passed through the `IIR` filter as $s^t = (1 - r)s^{t-1} + rs^t$ with the forgetting factor r. Alternatively, the scales could be updated individually for each particular part or segment, depending on a scenario. The parameters of (5) are then updated with the filtered scale as $\mu^t_{i,j;i,j+1} = s^t \mu^{t-1}_{i,j;i,j+1}$ and $\sigma^t_{i,j;i,j+1} = s^t \sigma^{t-1}_{i,j;i,j+1}$.

The updated orientation Θ^{t-1}_i in (6) will be the reference for $\theta_{i,j;i,j+1}$ in segment $i \in 1, \ldots, K$ in the next frame I^{t+1}. Knowing the inter-frame correspondences between the MAP locations of parts L^{t-1}_i and L^t_i (Fig. 3) of i-th segment, Θ^t_i is obtained through Kabsch algorithm [18] that estimates segment's rotation $\mathbf{R}^t_i$ in the least-squares sense by solving[1]:

$$\underset{\mathbf{R}^t_i}{\operatorname{argmin}} \sum_{j=1}^{k_i} \left\| \hat{l}^t_{i,j} - s^t \mathbf{R}^t_i \hat{l}^{t-1}_{i,j} \right\|^2_2 \tag{7}$$

Note, that the points $\hat{l}^t_{i,j}$ and $\hat{l}^{t-1}_{i,j}$ are translated to the origins of their respective CSs, with necessary rescaling of the latter. The stiffness parameters κ_i remain constant, as they are assumed invariant to any object deformations and change in viewpoint.

[1] For ease of readability, we drop the index $t-1$ in the notation of scale change and rotation from frame $t-1$ to t.

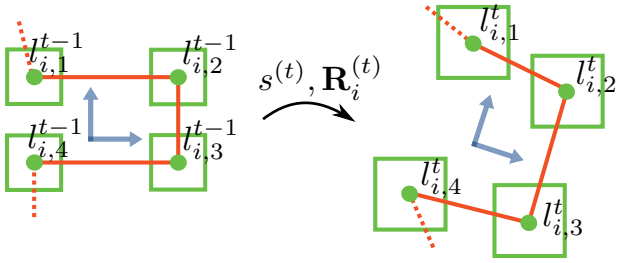

Figure 3. Synthetic example of i-th segment of heavily deformable object, whose scale s^t increases. The segment, consisting of $k_i = 4$ parts, deforms and rotates by $\mathbf{R}_i^t$ between two successive frames. The corresponding locations of parts between frames, translated back to the origin of the 2D CS, allow for recovering segment's rotation $\mathbf{R}_i^t$ despite its incident deformation. The dotted links connect it to neighbor segments, which rotate independently of the i-th segment.

4. Experimental results

In this Section, we experimentally challenge the *versatility* of our model (1). We show that our pose-configurable system can be used successfully to track elongated objects in the image plane, which can shift, rotate, change scale, be rigid and deform by flexing, articulating, and vibrating.

We also quantitatively evaluate our tracker on PROST dataset [34] with challenges of fast viewpoint changes, motion blur, heavy scale and illumination changes, and frequent occlusions. The tracker is compared against state-of-the-art trackers on PROST that learn their appearance online. We demonstrate that our spatio-temporal model with remarkably simple, fixed appearance term leads to competitive or better tracking performance. As the occlusion event is not modeled explicitly, we enforce constant appearance so that the tracker is robust against occlusions and thus can recover easily by redetecting the object.

Implementation details: In all experiments, we have the following, *fixed* settings. For (2), we use 512-dimensional RGB color histograms $h_{i,j}$ (8 bins per channel), with weighting $\nu_{i,j} = 2.0$. The mean distance in (5) is computed from the initial locations of the parts as $\mu^0_{i,j;i,j+1} = \left\| l^0_{i,j} - l^0_{i,j+1} \right\|_2$, while the standard deviation $\sigma^0_{i,j;i,j+1}$ as the average of their window radii that are close to Θ^0_i. The angular stiffness parameters κ_i of (6) correspond to $60°$. States of each part are defined over a regular, sparse 3D grid with size twice ($\times 2$) the size of the part. The scale is partitioned as $\times 0.9$, $\times 1.0$, $\times 1.1$ and filtered with $r = 0.1$.

Our tracking algorithm is a C++ single-threaded implementation (without SSE). It ran on a plain PC equipped with Intel Xeon@2.4 GHz, 4 MB cache, and 3.5 GB RAM. The frame processing speed scales linearly with the number of parts but also depends on their window sizes (optionally, the latter could be factored out with [30]).

Qualitative evaluation: We demonstrate that our method applies to tracking elongated objects of various shapes, which are rigid or deform by flexing, articulating, and vibrating in the image plane. The instances are tracked in the video sequences *Liquor* [34], *Surgical suture* [27], *Robotic arm*, *Toy tram*, *Guitar string*[2], shown in Fig. 4.

In *Liquor*, the tracker is very successful despite multiple and heavy occlusions of the glass bottle and is not confused by another bottle, which is fairly similar in color. In *Robotic arm*, the tracker follows the 2D pose of the articulating robotic manipulator composed of two segments. In *Surgical suture*, the suture is a very long object, which is thin and deforms heavily and unsystematically. By splitting the suture into piece-wise linear segments, our pose-configurable system can follow it very precisely. Despite no constraints at the ends of the suture, the tracker stabilized both ends correctly, which is a challenging task [15]. We posit this satisfactory behavior owes to the fact that, while some segments rotate, others only shift, and thus our hierarchical, spatio-temporal model renders the tracker stable. In *Toy tram*, our model can explain the bending and scale change of the tram and is robust against moderate out-of-plane rotations affecting its appearance. In *Guitar string*, the tracker is able to precisely register intricate deformations of the string with very little information available. The parts have only few pixels. In this case though, the tracker ran with fixed scale to prevent the model from shrinking on the textureless, string region. For comparison, the same sequence with scale update is shown in Fig. 5.

Quantitative evaluation: We evaluate quantitatively our approach on PROST. We can easily configure our region-based model to rigid objects with K=1 segment at initial orientation Θ^0_1, and partition it *evenly* into $k_1 = 3$ parts, i.e. such that the parts span the segment with no (or very small) overlap (see, *e.g.*, top rows in Fig. 4, 5). We then use the following evaluation measures:

- Intersection-over-union, as in [34],
- Mean distance precision, as in [5].

Specifically, the first criterion renders a detection as true positive when its bounding box overlaps with the ground truth bounding box by $> 50\%$. The recall performance is reported as the number of true positives over the sum of true positives and false negatives. The second criterion computes the ℓ_2-distance between the centers of detected and ground truth bounding-boxes.

To make the comparison fair, we fix the scale of our tracker and always output the same size of the ground truth bounding-box. Note that in the first frame of each sequence, our tracker outputs center location of the whole object that is slightly misaligned (by several pixels) from the center of

[2]Last 3 video sequences were collected from YouTube.

Figure 4. Our qualitative results on sequences (best viewed in color), enumerated from top to bottom. We display each example individually for better visualization. The left column shows initialized layouts of chained segments of *evenly* annotated parts. Their corresponding orientations, updated over time, are depicted on image sides together with frame number and frame rate. (i) The glass bottle is configured with $K = 1$ segment of $k_1 = 3$ parts. (ii) Articulating robotic arm is split into $K = 2$ segments of $k_1 = 6$ and $k_2 = 5$ parts. (iii) We split surgical suture into $K = 6$ segments of $k_i = 11$ parts. (iv) The tram only bends so we configure it with $K = 1$ segment of $k_1 = 5$ parts. (v) One can expect the vibrating string to deform only slightly, so we configure it with $K = 1$ segment, as well. We split it into 114 parts, as we observed that its registered motion was more realistic with the increased number of its "mini-parts".

the ground truth bounding-box, as it averages the locations of all its parts. For this reason, we precompute this misalignment vector in the first frame and fix it for the whole duration of the sequence. Then, in subsequent frames, the tracker shifts our center by the above, constant offset.

The quantitative results are shown in Table 1. Our tracker with constant appearance yields competitive performance with respect to TLD [19] and GD [22], while outperforming others, and processes videos at $\sim$ 100 `fps`. GD used scale update for evaluation though, while TLD struggles with deformable objects [19]. Our method performs best (top table) on the *Liquor* sequence with blur and multiple, partial and full occlusions and on the *Lemming* sequence with blur and heavy scale changes. Interestingly, it outperforms FT [1], which likewise splits an object into parts and fixes their appearance. For the sake of coherency of the experiments, we also ran our algorithm to detect scale change. We observed comparable performance (see, *e.g.*, top row in Fig. 4) and the frame rate was $\sim$ 45 `fps`. Our tracker was worse on the *Box* sequence (similarly to FT), in which the box drastically changed its appearance due to heavy illumination change on its reflective surface (Fig. 5). Hence, our model should benefit from additional features in the appearance term. This work aimed at a general method with the focus on the strong spatio-temporal and basic appearance models that jointly led to state-of-the-art results at very high frame rate.

5. Conclusions

This paper proposes a tracker that: (i) is modular, with one-shot learned spatio-temporal model and *without* dedicated model of dynamics, (ii) explicitly estimates various motions in the image plane, including deformations, by outputting detailed 2D pose (locations and scales of parts and orientations of segments), (iii) is robust against appearance changes resulting from change of viewpoint and occlusions, (iv) yields simple implementation with low computational cost allowing rates up to 100 `fps`, (v) scales efficiently from low- to high-dimensional state spaces, thus demonstrating that our single model can be easily *reconfigured* from one elongated object to another, (vi) uses a remarkably simple, fixed appearance term yet providing competitive state-of-the-art results on the challenging benchmark.

Figure 5. Our current limitations (best viewed in color). **Top:** Since we integrate color histograms into our appearance term, the tracker struggles with heavy illumination changes, present in the *Box* sequence (*e.g.*, frames #349, #353). **Bottom:** Unlike snakes models, the tracker is confused on textureless regions and shrinks when it updates scale. In *Guitar string*, it cannot discern between the correct and smaller scale of the parts of the guitar string (with the same configuration as in Fig. 4) and blindly rescales the pairwise distance constraints. Integrating other features into the appearance term (*e.g.*, optic flow [34] or gradients [25]) might correctly address these challenges.

Complementary to on-line appearance update algorithms, our future work will pursue development of *on-line reconfiguration* update mechanisms for updating object rigidity constraints over time. Since the proposed generic tracker allows for attributing local rigidity constraints over the spatio-temporal space occupied by various elongated objects, it thus opens opportunities to investigate dynamic adaptation of rigidity constraints for more robust tracking.

Acknowledgments. This work was partly supported by European FP7 project 216487 CompanionAble.

Method	Sequence				Avg
	Brd	Box	Lem.	Liq.	
Intersection-over-union[%] [34]					
MIF[41]	92.1	42.9	88.1	75.6	74.7
ORF[33]	10.0	28.3	17.2	53.6	27.3
FT[1]	67.9	61.4	54.9	79.9	66.0
MIL[5]	67.9	24.5	83.6	20.6	49.2
PROST[34]	75.0	90.6	70.5	85.4	80.4
GD[22]	**94.3**	**91.8**	78.0	91.4	88.9
TLD[19]	87.1	**91.8**	85.8	91.7	**89.1**
Ours	90.7	63.1	**91.4**	**96.6**	85.5
Mean distance precision [5]					
NN[13]	20.0	16.9	79.1	15.0	32.8
MIF[41]	13.7	63.7	19.4	42.5	34.8
ORF[33]	154.5	145.4	166.3	67.3	133.4
FT[1]	90.1	57.4	82.8	30.7	65.3
MIL[5]	51.2	104.6	14.9	165.1	84.0
PROST[34]	39.0	**13.0**	25.1	21.5	24.7
GD[22]	14.7	13.2	28.4	11.9	17.1
TLD[19]	**10.9**	17.4	16.4	6.5	**12.8**
Ours	18.9	43.2	**12.9**	**5.6**	20.2

Table 1. Performance of our configurable algorithm, evaluated on PROST database. For overall comparison, we also provide average scores in the rightmost column. Best results for each measure are indicated in **bold**.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *Computer Vision and Pattern Recognition*, pages 798–805, 2006. 2, 6, 7

[2] A. Amini, S. Tehrani, and T. Weymouth. Using dynamic programming for minimizing the energy of active contours in the presence of hard constraints. In *International Conference on Computer Vision*, pages 95–99, 1988. 3

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Computer Vision and Pattern Recognition*, pages 1014–1021, 2009. 2

[4] B. Babenko, M. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1619–1632, 2011. 2

[5] B. Babenko, M. Yang, and S. J. Belongie. Visual tracking with online Multiple Instance Learning. In *Computer Vision and Pattern Recognition*, pages 983–990, 2009. 1, 5, 7

[6] L. Cehovin, M. Kristan, and A. Leonardis. An adaptive coupled-layer visual model for robust visual tracking. In *International Conference on Computer Vision*, pages 1363–1370, 2011. 1, 2

[7] R. Collins. Mean-shift blob tracking through scale space. In *Computer Vision and Pattern Recognition*, pages 234–240, 2003. 2

[8] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition*, 2000. 2

[9] Z. Fan, M. Yang, and Y. Wu. Multiple collaborative kernel tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1268–1273, 2007. 2

[10] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. 2, 4

[11] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22(1):67 – 92, 1973. 2

[12] M. Godec, P. Roth, and H. Bischof. Hough-based tracking of non-rigid objects. In *International Conference on Computer Vision*, pages 81–88, 2011. 1, 2

[13] S. Gu, Y. Zheng, and C. Tomasi. Efficient visual object tracking with online nearest neighbor classifier. In *Asian Conference on Computer Vision*, pages 271–282, 2010. 7

[14] G. Hager, M. Dewan, and C. Stewart. Multiple kernel tracking with SSD. In *Computer Vision and Pattern Recognition*, pages 790–797, 2004. 2

[15] T. H. Heibel, B. Glocker, M. Groher, N. Paragios, N. Komodakis, and N. Navab. Discrete tracking of parametrized curves. In *Computer Vision and Pattern Recognition*, pages 1754–1761, 2009. 3, 5

[16] M. Isard and A. Blake. CONDENSATION - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998. 2

[17] H. Jiang, T.-P. Tian, K. He, and S. Sclaroff. Scale resilient, rotation invariant articulated object matching. In *Computer Vision and Pattern Recognition*, pages 143–150. IEEE, 2012. 2, 3

[18] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34(5):827–828, 1978. 4

[19] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1409–1422, 2012. 1, 6, 7

[20] L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The chains model for detecting parts by their context. In *Computer Vision and Pattern Recognition*, pages 25–32, 2010. 3

[21] M. Kass, A. Witkin, and D. Terzopoulos. Snakes - active contour models. *International Journal Of Computer Vision*, 1(4):321–331, 1987. 3

[22] D. Klein and A. Cremers. Boosting scalable gradient features for adaptive real-time tracking. In *International Conference on Robotics and Automation*, pages 4411–4416, 2011. 1, 6, 7

[23] J. Kwon and K. Lee. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In *Computer Vision and Pattern Recognition*, 2009. 1

[24] I. Leichter. Mean shift trackers with cross-bin metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):695–706, 2012. 2

[25] E. Maggio, F. Smeraldi, and A. Cavallaro. Combining colour and orientation for adaptive particle filter-based tracking. In *British Machine Vision Conference*, 2005. 3, 7

[26] T. Mauthner, M. Donoser, and H. Bischof. Robust tracking of spatial related components. In *International Conference on Pattern Recognition*, pages 1–4, 2008. 2

[27] N. Padoy and G. Hager. Deformable tracking of textured curvilinear objects. In *British Machine Vision Conference*, 2012. 5

[28] D. Park and D. Ramanan. N-best maximal decoders for part models. In *International Conference on Computer Vision*, pages 2627–2634, 2011. 1

[29] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *European Conference on Computer Vision*, volume 2350, pages 661–675, 2002. 2

[30] F. M. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *Computer Vision and Pattern Recognition*, pages 829–836, 2005. 5

[31] D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):65–81, 2007. 2, 4

[32] R. Ronfard. Region-based strategies for active contour models. *International Journal of Computer Vision*, 13(2):229–251, 1994. 3

[33] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof. On-line random forests. In *International Conference on Computer Vision Workshop on On-line Learning for Computer Vision*, 2009. 1, 7

[34] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof. PROST: Parallel robust online simple tracking. In *Computer Vision and Pattern Recognition*, pages 723–730, 2010. 1, 5, 7

[35] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *Computer Vision and Pattern Recognition*, 2011. 3

[36] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. 1

[37] S. Sclaroff and J. Isidoro. Active blobs: region-based, deformable appearance models. *Computer Vision and Image Understanding*, 89(2-3):197–225, 2003. 3

[38] S. M. N. Shahed, J. Ho, and M. Yang. Online visual tracking with histograms and articulating blocks. *Computer Vision and Image Understanding*, 114(8):901–914, 2010. 2

[39] C. Shen, M. J. Brooks, and A. van den Hengel. Fast global kernel density mode seeking with application to localisation and tracking. In *International Conference on Computer Vision*, pages 1516–1523, 2005. 2

[40] D. Wesierski, P. Horain, and Z. Kowalczuk. EBE: Elastic Blob Ensemble for coarse human tracking. In *International Conference on Image Processing*, pages 1625–1628, 2012. 3

[41] K. Wnuk and S. Soatto. Multiple instance filtering. In *Neural Information Processing Systems*, pages 370–378, 2011. 1, 7

[42] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixtures of parts. In *Computer Vision and Pattern Recognition*, 2011. 2, 4

[43] A. Yilmaz. Object tracking by asymmetric kernel mean shift with automatic scale and orientation selection. In *Computer Vision and Pattern Recognition*, 2007. 2, 4

[44] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition*, 2012. 2