

Deformable Parts Correlation Filters for Robust Visual Tracking

Alan Lukežič, Luka Čehovin, *Member, IEEE*, and Matej Kristan, *Member, IEEE*

Abstract—Deformable parts models show a great potential in tracking by principally addressing non-rigid object deformations and self occlusions, but according to recent benchmarks, they often lag behind the holistic approaches. The reason is that potentially large number of degrees of freedom have to be estimated for object localization and simplifications of the constellation topology are often assumed to make the inference tractable. We present a new formulation of the constellation model with correlation filters that treats the geometric and visual constraints within a single convex cost function and derive a highly efficient optimization for MAP inference of a fully-connected constellation. We propose a tracker that models the object at two levels of detail. The coarse level corresponds a root correlation filter and a novel color model for approximate object localization, while the mid-level representation is composed of the new deformable constellation of correlation filters that refine the object location. The resulting tracker is rigorously analyzed on a highly challenging OTB, VOT2014 and VOT2015 benchmarks, exhibits a state-of-the-art performance and runs in real-time.

Index Terms—Computer vision, visual object tracking, correlation filters, spring systems, short-term tracking.

1 INTRODUCTION

Short-term single-object visual tracking has received a significant attention of the computer vision community over the last decade with numerous conceptually diverse tracking algorithms being proposed every year. Recently several papers reporting experimental comparison of trackers on a common testing ground have been published [1], [2], [3], [4]. Results show that tracking quality depends highly on the expressiveness of the feature space in the object appearance model and the inference algorithm that converts the features into a presence score in the observed parameter space. Most of the popular trackers apply holistic appearance models which capture the object appearance by a single patch. In combination with efficient machine-learning and signal processing techniques from online classification and regression, these trackers exhibited top performance across all benchmarks [5], [6], [7], [8]. Most of these approaches apply sliding windows for object localization, and some extend the local search in the scale space [9], [10], [11], [12] to address the scale changes as well.

Nevertheless, a single patch often poorly approximates objects that undergo significant, potentially nonlinear, deformation, self occlusion and partial occlusions, leading to drift, model corruption and eventual failure. Such situations are conceptually better addressed by part-based models that decompose the object into a constellation of parts. This type of trackers shows a great potential in tracking non-rigid objects, but their performance often falls behind the holistic models [4], because of the large number of degrees of freedom that have to be estimated in the deformation model during tracking. Čehovin et al. [13] therefore propose

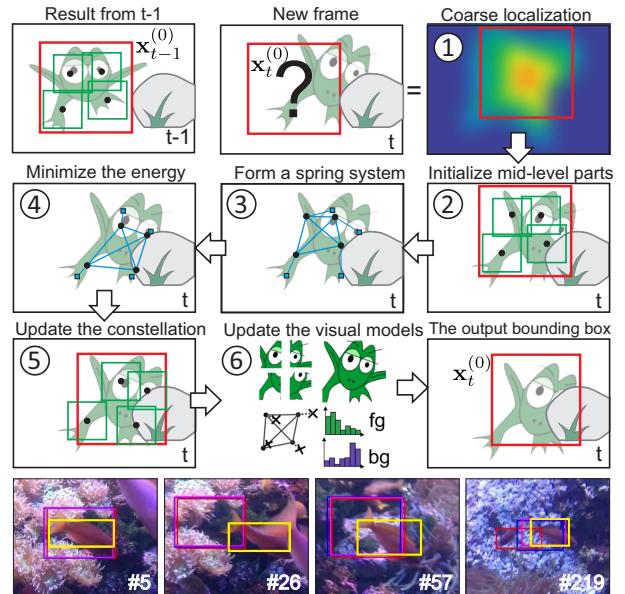


Fig. 1. Illustration of coarse-to-fine tracking by spring system energy minimization in a deformable part model (top). Tracking examples with our tracker DPT (yellow), KCF (red), IVT (blue) and Struck (magenta) are shown in the bottom.

that part-based models should be considered in a layered framework that decomposes the model into a global and local layer to increase the stability of deformation parameters estimation in presence of uncertain visual information. Most part-based trackers use very small parts, apply low-level features for the appearance models, e.g., histograms [13], [14] or keypoints [15], [16] and increase their discrimination power by increasing the number of parts. Object is localized by optimizing a trade-off between the visual and geometric agreement. Most of the recent trackers use star-based topo-

• A. Lukežič, L. Čehovin and M. Kristan are with the Faculty of Computer and Information Science, University of Ljubljana, Slovenia.
E-mail: see <http://www.vicos.si/>

ogy, e.g. [14], [15], [17], [18], [19], [20], or local connectivity, e.g. [13], instead of a fully-connected constellation [16] to make the inference tractable, but at a cost of a reduced power of the geometric model.

In this paper we present a new class of layered part-based trackers that apply a geometrically constrained constellation of local correlation filters [8], [11] for object localization. We introduce a new formulation of the constellation model that allows efficient optimization of a fully-connected constellation and adds only a negligible overhead to the tracking speed. Our part-based correlation filter formulation is cast in a layered part-based tracking framework [13] that decomposes the target model into a coarse layer and a local layer. A novel segmentation-based coarse model is introduced as well. Our tracker explicitly addresses the nonrigid deformations and (self-)occlusions, resulting in increased robustness compared to the recently proposed holistic correlation filters [11] as well as state-of-the-art part-based trackers.

1.1 Related work

Popular types of appearance models frequently used for tracking are generative holistic models like color histograms [21] and subspace-based [22], [23] or sparse reconstruction templates [24]. Several papers explored multiple generative model combinations [21], [25] and recently Gaussian process regressors were proposed for efficient updating of these models [26]. The cost function in generative holistic models reflects the quality of global object reconstruction in the chosen feature space, making the trackers prone to drifting in presence of local or partial object appearance changes or whenever the object moves on a visually-similar background. This issue is better addressed by the discriminative trackers which train an online object/background classifier and apply it to object localization. Early work includes support vector machines (SVM) [27], online Adaboost [7], multiple-instance learning [6] and recently excellent performance was demonstrated by structured SVMs [5]. A color-based discriminative model was recently presented in [28] that explicitly searches for potential visual distractors in the object vicinity and updates the model to increase the discriminative power. The recent revival of the matched filters [29] in the context of visual tracking has shown that efficient discriminative trackers can be designed by online learning of a correlation filter that minimizes the signal-to-noise ratio cost function. These filters exhibit excellent performance at high speeds, since learning and matching is carried out by exploiting the efficiency of the fast Fourier transform. Bolme et al. [8] introduced the first successful online matched filter, now commonly known as a correlation filter tracker. Their tracker was based on grayscale templates, but recently the correlation filters have been extended to multidimensional features [9], [10], [11], and Henriques et al. [11] introduced kernelized versions. Scale adaptation of correlation filters was investigated by Danneljan et al. [9] and Zhang et al. [30] who applied correlation filters to the scale space and [31] who combined votes of multiple automatically allocated filters. Zhang et al. [32] have shown the connection to spatio-temporal context learning. Hong et al. [33] have recently integrated correlation filters in a multi-store tracking framework and demonstrated

excellent performance. In fact, the correlation filter-based trackers have demonstrated excellent performance across all the recent benchmarks. Still, these trackers suffer from the general drawbacks of holistic models is that they do not explicitly account for deformation, self occlusion and partial occlusions, leading to drift, model corruption and eventual failure. This issue is conceptually better addressed by models that decompose the object into parts.

The part-based trackers apply constellations of either generative or discriminative local models and vary significantly in the way they model the constellation geometry. Hoey [34] used a flock-of-features tracking in which parts are independently tracked by optical flow. The flock is kept on object by identifying parts that deviate too far from the flock and replacing them with new ones. But because of weak geometric constraints, tracking is prone to drifting. Vojir et al. [35] addressed this issue by significantly constraining the extent of each part displacement and introduced tests of estimation quality. Tracking robustness is increased by only considering the part displacements deemed accurately estimated. Martinez et al. [36] proposed connecting triplets of parts and tracked them by kernels while enforcing locally-affine deformations. The local connectivity resulted in inefficient optimization and parts required careful manual initialization. Artner et al. [16] proposed a key-point-based tracker with a fully-connected constellation. They use the geometric model that enforces preservation of interkeypoint distance ratios. Because the ratios are not updated during tracking and due to the ad-hoc combination of geometric and appearance models, the resulting optimization is quite brittle, requiring manual initialization of parts and the resulting tracker handles only moderate locally-affine deformations. Pernici et al. [37] address nonrigid deformations by oversampling key-points to construct multiple instance-models and use a similarity transform for matching. But, the tracker still fails at significant nonrigid deformations. Several works simplify a geometric model to a star-based topology in interest of simplified optimization. A number of these works apply part detectors and a generalized Hough transform for localization. Examples of part detectors are key-points [15], random forest classifiers [19], ferns [38] and pixels [39]. Cai et al. [17] apply superpixels as parts combined with segmentation for efficient tracking, but the high reliability on color results in significant failures during illumination changes. Kwon et al. [14] apply generative models in a star-based topology with adding and removing parts and Čehovin et al. [13] increase the power of the geometric model by local connectivity. Both approaches require efficient stochastic optimizers for inference. Yao et al. [18] address the visual and geometric model within a single discriminative framework. They extend the structured SVM [5] to multiple part tracking, but cannot handle scale changes. This model was extended by Zhu et al. [20] to account for context as well, but uses a star-based topology for making the inference tractable. Context was also used by Duan et al. [40] where tracking multiple objects or object parts was used to resolve ambiguities.

Part-based trackers often suffer from the potentially large number of parameters of the deformation model to be estimated from uncertain/noisy visual data. This is addressed by the layered paradigm of part-based trackers in-

troduced by Čehovin et al. [13]. This paradigm decomposes the tracker architecture into a global coarse and a local appearance layer. The global layer contains coarse target representations such as holistic templates and global color histograms, while the local layer is the constellation of parts with simple local appearance description. The paradigm applies a top-down localization to gradually estimate the state parameters (i.e., target center and part locations) and bottom-up updates to update the appearance models. Čehovin et al. [13] analyzed various modalities used at the global layer (i.e., color, local motion and shape) and their influence on tracking. They have concluded that color plays the most important role at the scale of the entire object.

1.2 Our approach and contributions

Our main contribution is a new class of fully-connected part-based correlation filter trackers. Most part-based trackers apply star-based topology to simplify the inference or combine geometrical and visual constraints in an ad-hoc fashion often leading to a nonconvex optimization problem. In contrast, our formulation treats the geometric and visual constraints within a single convex cost function. We show that this cost function has a dual formulation of a spring system and show that MAP inference of the constellation can be achieved by minimizing the energy of the dual spring system. We derive a highly efficient optimizer that in practice results in a very small computational overhead during tracking.

The tracker is formulated within the theoretical framework of layered deformable parts [13] that decomposes the tracker into a coarse representation and a mid-level representation. The coarse representation is composed of a holistic correlation filter and a novel global color model. The mid-level representation is composed of local correlation filters fully-connected by the new constellation model. Tracking is performed by top-down localization and bottom-up updates (Figure 1): The coarse model initializes the mid-level representation at approximate object location. An equivalent spring system is formed and optimized, yielding a MAP constellation estimate. The parts are updated and the estimated constellation is used to update the coarse model. In contrast to the standard holistic correlation filters, the proposed deformable parts tracker naturally addresses the object appearance changes resulting from scale change, nonrigid deformations and (self)occlusions increasing the tracking robustness.

Our tracker and the proposed constellation optimization are analyzed in depth. The tracker is rigorously compared against a set of state-of-the-art trackers on a highly challenging recent benchmarks OTB [1], VOT2014 [4] and VOT2015 [41] and exhibits a state-of-the-art performance. Additional tests show that improvements come from the fully-connected constellation and the top-down/bottom-up combination of the coarse representation with the proposed deformable parts model.

2 DEFORMABLE PARTS TRACKER

As it is a common practice in visual tracking, the tracker output at time-step t is an axis-aligned bounding box.

In our case this region is estimated by the deformable parts correlation filter as we describe in this Section. Our tracker is composed of a coarse representation described in Section 2.2, and of a deformable constellation of parts, a mid-level object representation described in Section 2.3. In the following, we will denote the part positions by $(\cdot)^{(i)}$, where the index $i = 0$ denotes the root part in the coarse layer and indexes $i > 0$ denote parts in the constellation. Since both representations apply kernelized correlation filters (KCF) [11] for part localization, we start by briefly describing the KCF in Section 2.1.

2.1 Kernelized correlation filters

This section summarizes the main results of the recent advances in correlation filters and their application to tracking [11], [42]. Given a single grayscale image patch \mathbf{z} of size $M \times N$ a linear regression function $f(\mathbf{z}) = \mathbf{w}^T \mathbf{z}$ is estimated such that its response is maximal at the center of the patch and gradually reduces for the patch circular shifts $\mathbf{z}_{m,n}, (m, n) \in \{0, \dots, M-1\} \times \{0, \dots, N-1\}$ toward the patch edge. This is formulated by minimizing the following cost function

$$\epsilon = \|\mathbf{w} \otimes \mathbf{z} - \phi\|^2 + \lambda \|\mathbf{w}\|^2, \quad (1)$$

where \otimes denotes circular correlation, ϕ is a Gaussian function centered at zero shift (see Figure 2) and λ is a ridge regression regularization parameter which controls overfitting. The correlation in (1) is kernelized [11] by redefining

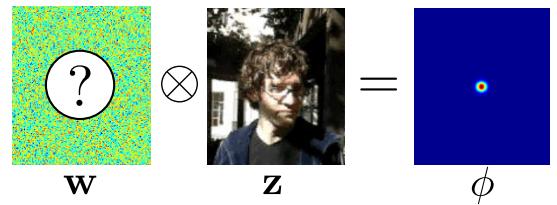


Fig. 2. The correlation filter formulation. We seek a weight matrix \mathbf{w} that results in a Gaussian response function ϕ when correlated over the image patch \mathbf{z} .

the \mathbf{w} as a linear combination of the circular shifts, i.e., $\mathbf{w} = \sum_{m,n} a_{m,n} \varphi(\mathbf{z}_{m,n})$, where $\varphi(\cdot)$ is a mapping to the Hilbert space induced by a kernel $\kappa(\cdot, \cdot)$. The minimum of (1) is obtained at

$$\mathbf{A} = \frac{\Phi}{U_z + \lambda}, \quad (2)$$

where the capital letters denote the Fourier transforms of image-domain variables, i.e., $\mathbf{A} = \mathcal{F}[\mathbf{a}]$, $\Phi = \mathcal{F}[\phi]$, $U_z = \mathcal{F}[\mathbf{u}_z]$, with $\mathbf{u}_z(m, n) = \kappa(\mathbf{z}_{m,n}, \mathbf{z})$ and \mathbf{a} is a dual representation of \mathbf{w} [11]. At time-step t , a patch \mathbf{y}_t of size $M \times N$ is extracted from the image and the probability of object at pixel location \mathbf{x}_t is calculated from the current estimate of \mathbf{A}_t and the template \mathbf{z}_t as

$$p(\mathbf{y}_t | \mathbf{x}_t, \mathbf{z}_t) \propto \mathcal{F}^{-1}[\mathbf{A}_t \odot \mathbf{U}_y], \quad (3)$$

where $\mathbf{U}_y = \mathcal{F}[\mathbf{u}_y]$, $\mathbf{u}_y(m, n) = \kappa(\mathbf{y}_{m,n}, \mathbf{z}_t)$. In [11], [42], the maximum on $p(\mathbf{y}_t | \mathbf{x}_t, \mathbf{z}_t)$ is taken as the new object position. The numerator and denominator of \mathbf{A}_t in (2) as well as the patch template \mathbf{z}_t are updated separately at

the estimated position by an autoregressive model. The extension of the kernelized filter from grayscale patches to multi-channel features is straigth-forward and we refer the reader to [11], [42] for details.

2.2 The coarse representation

The coarse object representation in our appearance model consists of two high-level object models: the object global template $\mathbf{z}_t^{(0)}$ (a root correlation filter) and a global color model $C_t = \{p(\mathbf{x}_t|f), p(\mathbf{x}_t|b)\}$, specified by the foreground and background color histograms, $p(\mathbf{x}_t|f)$ and $p(\mathbf{x}_t|b)$, respectively, where \mathbf{x}_t denotes the pixel coordinates. These models are used in each tracking iteration to coarsely estimate the center $\mathbf{x}_t^{(0)}$ of the object bounding box within a specified search region (Figure 1, step 1), which is subsequently refined by the mid-level representation (Section 2.3).

Given an image patch $\mathbf{y}_t^{(0)}$ extracted from a search region, (Figure 3a), the center is estimated by maximizing the probability of object location $\mathbf{x}_t^{(0)}$,

$$p(\mathbf{x}_t^{(0)}|\mathbf{z}_t^{(0)}, C_t, \mathbf{y}_t^{(0)}) \propto p(\mathbf{y}_t^{(0)}|\mathbf{x}_t^{(0)}, \mathbf{z}_t^{(0)})p(\mathbf{y}_t^{(0)}|\mathbf{x}_t^{(0)}, C_t). \quad (4)$$

The first term, $p(\mathbf{y}_t^{(0)}|\mathbf{x}_t^{(0)}, \mathbf{z}_t^{(0)})$, is the template probability reflecting the similarity between the patch centered at $\mathbf{x}_t^{(0)}$ and the object template $\mathbf{z}_t^{(0)}$ calculated as the response from the correlation filter (3), (see Figure 3b). The second term is the color probability defined as

$$p(\mathbf{y}_t^{(0)}|\mathbf{x}_t^{(0)}, C_t) = p(f|\mathbf{x}_t^{(0)}, \mathbf{y}_t^{(0)})(1 - \alpha_{\text{col}}) + \alpha_{\text{col}}, \quad (5)$$

where $p(f|\mathbf{x}_t^{(0)}, \mathbf{y}_t^{(0)})$ is the probability of a pixel at location $\mathbf{x}_t^{(0)}$ belonging to a foreground and α_{col} is a weak uniform distribution that addresses sudden changes of the object color, since the $p(f|\mathbf{x}_t^{(0)}, \mathbf{y}_t^{(0)})$ might be uninformative in these situations and would deteriorate localization. The value of α_{col} varies with a color informativeness as detailed in Section 2.2.1. The probability $p(f|\mathbf{x}_t^{(0)}, \mathbf{y}_t^{(0)})$ is calculated by histogram backprojection, i.e., by applying the Bayes rule with $p(\mathbf{x}_t|f)$ and $p(\mathbf{x}_t|b)$, and regularized by a Markov random field [43], [44] to arrive at a smoothed foreground posterior (Figure 3c). Multiplying the template and color probabilities yields the density $p(\mathbf{x}_t^{(0)}|\mathbf{z}_t^{(0)}, C_t, \mathbf{y}_t^{(0)})$ (Figure 3d). Notice that on their own, the template and color result in ambiguous densities but their combination drastically reduces the ambiguity.

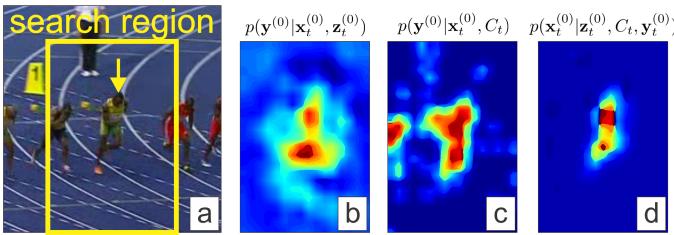


Fig. 3. Example of a search region and the tracked object indicated by a rectangle and an arrow (a). The coarse template probability, the color probability and the full coarse model density are shown in (b), (c) and (d), respectively.

2.2.1 Color informativeness test

Whenever the object color is similar to the background, or during sudden illumination variations, the color segmentation becomes unreliable and can degrade tracking performance. The color informativeness test is performed by comparing the number of pixels, $M_t^{(\text{fg})}$, assigned to the foreground by the color model $p(f|\mathbf{x}_t^{(0)}, \mathbf{y}_t^{(0)})$, and the object size from the previous time-step $M_{t-1}^{(\text{siz})}$ (i.e., the area of object bounding box). If the deviation from the expected object area is within the allowed bounds, the uniform component in (5) is set to a low value, otherwise it is set to 1, effectively ignoring the color information in the object position posterior (4), i.e.,

$$\alpha_{\text{col}} = \begin{cases} 0.1 & ; \alpha_{\text{min}} < \frac{M_t^{(\text{fg})}}{M_{t-1}^{(\text{siz})}} < \alpha_{\text{max}} \\ 1 & ; \text{otherwise} \end{cases} \quad (6)$$

The parameters α_{min} and α_{max} specify the interval of expected number of pixels assigned to the target relative to the target bounding box size from the previous time-step. Since the aim of (6) is only to detect drastic segmentation failures, these values can be set to a very low and very large value, respectively. Figure 4 illustrates the color informativeness test. In Figure 4(a), the number of pixels assigned to the foreground is within the expected bounds, while (b,c) show examples that fail the test by assigning too many or too few pixels to the object.

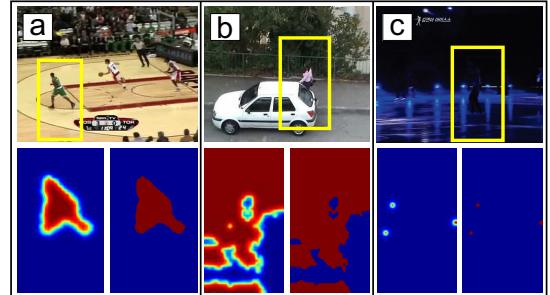


Fig. 4. Three examples of the color backprojection within the image patch denoted with the yellow bounding box. The regularized backprojection is shown on left and the binarized segmentation on right under each image. Example (a) passes the color informativeness test, while (b) and (c) fail the test since too many or too few pixels are assigned to the object.

2.3 The mid-level representation

The mid-level representation in our tracker is a geometrically constrained constellation of N_p parts $\mathbf{X}_t = \{\mathbf{x}_t^{(i)}\}_{i=1:N_p}$, where $\mathbf{x}_t^{(i)}$ is the position of i -th part (see Figure 5, left). Note that the part sizes do not change during tracking and therefore do not enter the state variable $\mathbf{x}_t^{(i)}$. Each part centered at $\mathbf{x}_t^{(i)}$ is a local mid-level representation of object, a kernelized correlation filter, specified by a fixed-size part template $\mathbf{z}_t^{(i)}$ and $\mathbf{A}_t^{(i)}$ (Section 2.1).

The probability of the constellation being at state \mathbf{X}_t conditioned on the parts measurements $\mathbf{Y}_t = \{\mathbf{y}_t^{(i)}\}_{i=1:N_p}$ and parameters of the deformation model Θ is decomposed into

$$p(\mathbf{X}_t|\mathbf{Y}_t, \Theta) \propto p(\mathbf{Y}_t|\mathbf{X}_t, \Theta)p(\mathbf{X}_t|\Theta). \quad (7)$$

The density $p(\mathbf{Y}_t | \mathbf{X}_t, \Theta)$ is the *measurement constraint* term, reflecting the agreement of measurements with the current state \mathbf{X}_t of constellation, whereas the second term, $p(\mathbf{X}_t | \Theta)$, reflects the agreement of the constellation with the *geometric constraints*.

2.3.1 Geometric constraints

The constellation is specified by a set of links $(i, j) \in \mathcal{L}$ indexing the connected pairs of parts (Figure 5). The parts and links form an undirected graph and the joint pdf over the part states can be factored over the links as

$$p(\mathbf{X}_t | \Theta) = \prod_{(i,j) \in \mathcal{L}} \phi(||d_t^{(i,j)}||; \mu^{(i,j)}, k^{(i,j)}), \quad (8)$$

where $d_t^{(i,j)} = \mathbf{x}_t^{(i)} - \mathbf{x}_t^{(j)}$ is a difference in positions of the linked parts, $\mu^{(i,j)}$ is the preferred distance between the pair of parts and $k^{(i,j)}$ is the intensity of this constraint. The factors in (8) are defined as Gaussians $\phi(\cdot; \mu, k)$ with mean μ and variance k meaning that deviations from the preferred distances decrease the probability (8).

2.3.2 Measurement constraints

Given a fixed part state, $\mathbf{x}_t^{(i)}$, the measurement $\mathbf{y}_t^{(i)}$ at that part is independent from the states of other parts. The measurement probability decomposes into a product of per-part visual likelihoods

$$p(\mathbf{Y}_t | \mathbf{X}_t, \Theta) = \prod_{i=1:N_p} p(\mathbf{y}_t^{(i)} | \mathbf{x}_t^{(i)}, \Theta). \quad (9)$$

To simplify the combination of the geometric and the visual constraints (Section 2.3.3) it is beneficial to chose the visual likelihoods from the same class of functions as (8). We make use of the fact that the parts appearance models are correlation filters trained on Gaussian outputs, thus the visual likelihoods in (9) can be defined as Gaussians as well. Let $\mathbf{x}_{tA}^{(i)}$ be the position in vicinity of $\mathbf{x}_t^{(i)}$ that maximizes the similarity of the appearance model $\mathbf{z}_t^{(i)}$ and the measurement $\mathbf{y}_t^{(i)}$ (see Figure 5, left). The visual likelihood can then be defined as a Gaussian $p(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}, \Theta) = \phi(||d_t^{(i)}||; 0, k^{(i)})$ where $d_t^{(i)} = \mathbf{x}_t^{(i)} - \mathbf{x}_{tA}^{(i)}$ is the difference of the part current state and its visually-ideal position, and $k^{(i)}$ is the intensity of this constraint.

2.3.3 The dual spring-system formulation

Substituting equations (8,9) back into (7) leads to an exponential posterior $p(\mathbf{X}_t | \mathbf{Y}_t, \Theta) \propto \exp(-E)$, with

$$E = \frac{1}{2} \sum_{i=1:N_p} k_t^{(i)} \|d_t^{(i)}\|^2 + \sum_{i,j \in \mathcal{L}} k_t^{(i,j)} (\mu_t^{(i,j)} - \|d_t^{(i,j)}\|)^2. \quad (10)$$

Note that E corresponds to an energy of a spring system in which pairs of parts are connected by springs and each part is connected by another spring to an image position most similar to the part appearance model (Figure 5, right). The terms $\mu^{(i,j)}$ and $k^{(i,j)}$ are nominal lengths and stiffness of springs interconnecting parts (dynamic springs), while $k^{(i)}$ is stiffness of the spring connecting part to the image location (static spring). In the following we will refer to the nodes in the spring system that correspond to parts that move during optimization as *dynamic nodes* and we will refer

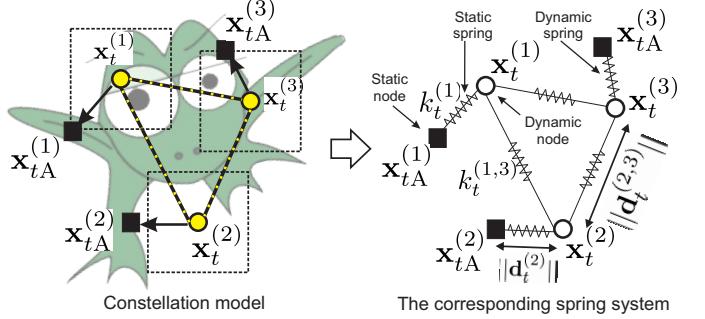


Fig. 5. Example of a constellation model with rectangular parts and arrows pointing to the most visually similar positions (left) and the dual form corresponding to a spring system (right). A constellation with only three nodes is shown for clarity.

to the nodes that are anchored to image positions as *static nodes*, since they do not move during the optimization.

The stiffness $k_t^{(i)}$ of a spring connecting a part to the image (in Figure 5 denoted as *static spring*) should reflect the uncertainty of the visually best-matching location $\mathbf{x}_{tA}^{(i)}$ in the search region of the i -th part and is set by the output of the correlation filter. The best matching position $\mathbf{x}_{tA}^{(i)}$ is estimated as location at which the output of the corresponding correlation filter (3) reaches a maximum value (denoted as $w_t^{(i)}$) and the spatial uncertainty in the search region is estimated as the weighted variance $\sigma_t^{2(i)}$, i.e., the average of squared distances from $\mathbf{x}_{tA}^{(i)}$ weighted by the correlation filter response map. The spring stiffness is thus defined by the response strength $w_t^{(i)}$ and spatial uncertainty, i.e.,

$$k_t^{(i)} = w_t^{(i)} / \sigma_t^{2(i)}. \quad (11)$$

The stiffness of springs interconnecting the parts (in Figure 5 denoted as *dynamic spring*) should counter significant deviations from the spring nominal length. Let $d_{tA}^{(i,j)} = \mathbf{x}_{tA}^{(i)} - \mathbf{x}_{tA}^{(j)}$ be the position difference between the visually most similar positions of the nodes indexed by i and j . The stiffness of the spring connecting the nodes is set to

$$k_t^{(i,j)} = \left(\frac{\mu_{t-1}^{(i,j)} - \|d_{tA}^{(i,j)}\|}{\mu_{t-1}^{(i,j)}} \right)^2. \quad (12)$$

2.4 Efficient MAP inference

The spring system from Section 2.3.3 is a dual representation of the deformable parts model and minimization of its (convex) energy function (10) corresponds to the maximum a posteriori state estimation (7) of the deformable parts model. This means that general-purpose convex energy minimizers can be used to infer the MAP state. But due to the dual spring system formulation, even more efficient optimizers can be derived. In particular, we propose an algorithm that splits a 2D spring system into two 1D systems, solves each in a closed form and then re-assembles them back into a 2D system (see Figure 6). This partial minimization is iterated until convergence. In the following we derive an efficient closed-form solver for a 1D system.

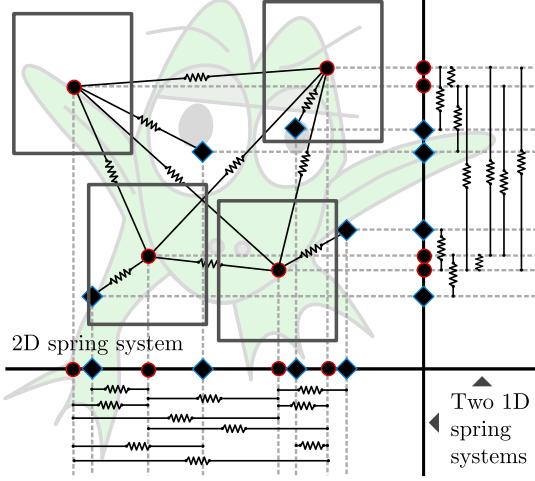


Fig. 6. Example of decomposition of a 2D spring system with 4 dynamic nodes (circles) and 4 static nodes (diamonds) on two 1D spring systems. Each 1D spring system has a closed-form solution.

Using standard results from Newtonian mechanics, the forces at springs F of a 1D spring system, can be written as

$$\mathbf{F} = -\mathbf{K}(\mathbf{Bx} - \mathbf{L}), \quad (13)$$

where $\mathbf{K} = \text{diag}([k_1, \dots, k_N])$ is a diagonal matrix of spring stiffness coefficients, \mathbf{x} is a vector of 1D nodes positions, $\mathbf{L} = [l_1, \dots, l_N]$ is a vector of spring nominal lengths and \mathbf{B} is a $N_{\text{springs}} \times N_{\text{nodes}}$ connectivity matrix that represents directed connections between the nodes. Let $\{n_{i1}, n_{i2}\}$ be indexes of two nodes connected by the i -th spring. The entries of \mathbf{B} are then defined as

$$b_{ij} = \begin{cases} 1 & ; j \equiv n_{i1} \\ -1 & ; j \equiv n_{i2} \\ 0 & ; \text{otherwise} \end{cases} \quad (14)$$

The forces at nodes $\mathbf{F}_{\text{nodes}}$ are given by left-multiplication of (13) by \mathbf{B}^T , yielding

$$\mathbf{F}_{\text{nodes}} = -\mathbf{B}^T \mathbf{K} \mathbf{B} \mathbf{x} + \mathbf{B}^T \mathbf{K} \mathbf{L}. \quad (15)$$

The equilibrium is reached when the forces at nodes vanish (i.e., become zero), resulting in the following linear system

$$\hat{\mathbf{K}} \mathbf{x} = \mathbf{C} \mathbf{L}, \quad (16)$$

where $\hat{\mathbf{K}} = \mathbf{B}^T \mathbf{K} \mathbf{B}$ and $\mathbf{C} = \mathbf{B}^T \mathbf{K}$. We will assume the following ordering in the nodes positions vector, $\mathbf{x} = [\mathbf{x}_{\text{dyn}}, \mathbf{x}_{\text{stat}}]^T$, where \mathbf{x}_{dyn} and \mathbf{x}_{stat} are 1D positions of the dynamic and static nodes, respectively. The matrix $\hat{\mathbf{K}}$ can be written as

$$\hat{\mathbf{K}} = \begin{bmatrix} \hat{\mathbf{K}}_{\text{dyn}} & \hat{\mathbf{K}}_{\text{stat}} \\ \hat{\mathbf{K}}_{\text{rem}} & \end{bmatrix}, \quad (17)$$

where $\hat{\mathbf{K}}_{\text{dyn}}$ and $\hat{\mathbf{K}}_{\text{stat}}$ are $N_{\text{dyn}} \times N_{\text{dyn}}$ and $N_{\text{dyn}} \times N_{\text{stat}}$ submatrices, respectively, relating the dynamic nodes to each other and the static nodes. Similar decomposition can be performed on \mathbf{C} ,

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{\text{dyn}} \\ \mathbf{C}_{\text{stat}} \end{bmatrix}. \quad (18)$$

Substituting the definitions (17) and (18) into (16) yields the following closed form for the dynamic nodes positions \mathbf{x}_{dyn} ,

$$\mathbf{x}_{\text{dyn}} = \hat{\mathbf{K}}_{\text{dyn}}^{-1} (\mathbf{C}_{\text{dyn}} \mathbf{L} - \hat{\mathbf{K}}_{\text{stat}} \mathbf{x}_{\text{stat}}). \quad (19)$$

The optimization of a 2D spring system, which we call iterative direct approach (IDA), is summarized in the Algorithm 1. At each iteration, a 2D system is decomposed into separate 1D systems, each system is solved by (19) and the 2D system is re-assembled. The process is iterated until convergence. Note that $\hat{\mathbf{K}}_{\text{stat}} \mathbf{x}_{\text{stat}}$ and $\hat{\mathbf{K}}_{\text{dyn}}^{-1}$ can be calculated only once and remain unchanged during the optimization.

Algorithm 1 : Optimization of a 2D spring system.

Require:

Positions of dynamic and static nodes, \mathbf{x}_{dyn} and \mathbf{x}_{stat} , stiffness vector \mathbf{k} and adjacency matrix \mathbf{B} .

Ensure:

Equilibrium positions of dynamic nodes \mathbf{x}_{dyn} .

Procedure:

- 1: For each dimension separately construct $\hat{\mathbf{K}}_{\text{dyn}}$, $\hat{\mathbf{K}}_{\text{stat}}$ and \mathbf{C}_{dyn} according to (17) and (18).
 - 2: **while** stop condition **do**
 - 3: For each dimension do:
 - 4: * Extract 1D positions of dynamic nodes from \mathbf{x}_{dyn} .
 - 5: * Calculate the current 1D spring lengths vector \mathbf{L} .
 - 6: * Estimate new values of \mathbf{x}_{dyn} by solving (19).
 - 7: Reassemble the 2D system.
 - 8: **end while**
-

2.5 Deformable parts tracker (DPT)

The coarse representation and the mid-level constellation of parts from Section 2.2 and Section 2.3 are integrated into a tracker that localizes the object at each time-step within a search region by a top-down localization and bottom-up updates. In the following we will call this tracker a deformable parts correlation filter tracker and denote it by DPT for short. The tracker steps are visualized in Figure 1 and detailed in the following subsections.

2.5.1 Top-down localization

The object is coarsely localized within a search region corresponding to the root correlation filter centered at the object position from the previous time-step $t - 1$. The object center at time-step t is approximated by position that maximizes the conditional probability $p(\mathbf{x}_t^{(0)} | \mathbf{z}_t^{(0)}, C_t, \mathbf{y}_t^{(0)})$ from Section 2.2 and a coarse center translation from $t - 1$ to t is estimated (Figure 1, step 1). The mid-level representation, i.e, constellation of parts, is initialized by this translation. For each translated part $\mathbf{x}_t^{(i)}$, the part correlation filter is applied to determine the position of the maximum similarity response, $\mathbf{x}_{tA}^{(i)}$, along with the stiffness coefficients $k_t^{(i)}$ and $k_t^{(i,j)}$ as detailed in Section 2.3.3. A MAP constellation estimate $\hat{\mathbf{X}}_t$ is obtained by minimizing the energy (10) of the equivalent spring system optimization from Section 2.4 (Figure 1, steps 2-4).

2.5.2 Bottom-up update

The mid-level and coarse representations are updated as follows (Figure 1, steps 5,6). The part correlation filters and their appearance models $\mathbf{z}_t^{(i)}$ are updated at MAP estimates of part positions $\hat{\mathbf{x}}_t^{(i)}$. Updating all appearance models at constant rate might lead to drifting and failure whenever the object is partially occluded or self-occluded. An effective mechanism is applied to address this issue. A part is updated only if its response at the MAP position $\hat{\mathbf{x}}_t^{(i)}$ is at least half of the strongest response among all parts and if at least twenty percent of all pixels within the part region correspond to the object according to the segmentation mask estimated at the root part (Section 2.2). The nominal spring lengths (the preferred distances between parts) are updated by an autoregressive scheme

$$\mu_t^{(i,j)} = \mu_{t-1}^{(i,j)}(1 - \alpha_{\text{spr}}) + \|\hat{d}_t^{(i,j)}\| \alpha_{\text{spr}}, \quad (20)$$

where $\|\hat{d}_t^{(i,j)}\|$ is the distance between the parts (i, j) in the MAP estimate $\hat{\mathbf{X}}_t$ and α_{spr} is the update factor.

The coarse representation is updated next. The MAP object bounding box is estimated by $\hat{\mathbf{x}}_t^{(0)} = \mathbf{T}_t \hat{\mathbf{x}}_{t-1}^{(0)}$, where \mathbf{T}_t is a Euclidean transform estimated by least squares from the constellation MAP estimates $\hat{\mathbf{X}}_{t-1}$ and $\hat{\mathbf{X}}_t$. The root correlation filter $\mathbf{z}_t^{(0)}$ and the histograms in the global color model C_t are updated at $\hat{\mathbf{x}}_t^{(0)}$. A histogram $\mathbf{h}_t^{(f)}$ is extracted from $\hat{\mathbf{x}}_t^{(0)}$ and another histogram $\mathbf{h}_t^{(b)}$ is extracted from the search region surrounding $\hat{\mathbf{x}}_t^{(0)}$ increased by a factor α_{sur} . The foreground and background histograms are updated by an autoregressive model, i.e.,

$$p(\mathbf{x}_t | \cdot) = p(\mathbf{x}_{t-1} | \cdot)(1 - \alpha_{\text{hist}}) + \mathbf{h}_t^{(\cdot)} \alpha_{\text{hist}}, \quad (21)$$

where α_{hist} is the forgetting factor. To increase adaptation robustness, the histograms are not updated if the color segmentation fails the color informativeness test from Section 2.2.1. The top-down localization and bottom-up update steps are summarized in Algorithm 2.

2.5.3 Tracker initialization

The coarse representation at time-step $t = 1$ is initialized from the initial bounding box $\mathbf{x}_1^{(0)}$. The mid-level the constellation of parts is initialized by splitting the initial object bounding box into four equal non-overlapping parts. The part appearance models are initialized at these locations and the preferred distances between parts are calculated from the initialized positions.

3 EXPERIMENTAL ANALYSIS

This section reports experimental analysis of the proposed DPT. The implementation details are given in Section 3.1, Section 3.2 details the analysis of the design choices, Section 3.3 reports comparison to the related state-of-the-art, Section 3.4 reports performance on recent benchmarks and Section 3.5 provides qualitative analysis.

Algorithm 2 : A tracking iteration of a deformable parts correlation filter tracker.

Require:

Coarse model $\{\mathbf{x}_{t-1}^{(0)}, \mathbf{z}_{t-1}^{(0)}, \mathbf{C}_{t-1}\}$ and mid-level model $\{\mathbf{X}_{t-1}, \mathbf{Z}_{t-1}\}$ at time-step $t - 1$.

Ensure:

Coarse model $\{\mathbf{x}_t^{(0)}, \mathbf{z}_t^{(0)}, \mathbf{C}_t\}$ and mid-level model $\{\mathbf{X}_t, \mathbf{Z}_t\}$ at time-step t .

Procedure:

- 1: Coarsely estimate the object position by the root node (Section 2.2) and displace the mid-level parts.
 - 2: Calculate the part correlation filter responses and form a spring system according to Section 2.3.3.
 - 3: Estimate the MAP mid-level parts constellation by optimizing the energy of a dual spring system (Section 2.4).
 - 4: Update the root node position and size by the Euclidean transform fitted to the parts positions before and after MAP inference (Section 2.5.1).
 - 5: Update the spring system parameters and the constellation appearance models (Section 2.5.2).
 - 6: Update the coarse color model \mathbf{C}_t and correlation filter $\mathbf{z}_t^{(0)}$.
-

3.1 Implementation details and parameters

Our implementation uses a kernelized correlation filters (KCF) [11] with HOG [45] features and grayscale template in the part appearance models. All filter parameters and learning rate are the same as in [11]. The parts have to be large enough to capture locally visually-distinctive regions on the object and have to cover the object without significantly overlapping with each other. The size of the tracked targets therefore places a constraint on the maximal number of parts since their size reduces with this number. For small parts, the HoG features become unreliable. But even more pressing is the issue that the capture range of correlation filters is constrained by the template size and is even reduced in practice due to the effects of circular correlation used for learning and matching. Therefore, small parts increasingly lose the ability to detect large displacements. The parts have to be large enough to capture the object partial appearance at sufficient level of detail, therefore we set the number of parts to $N_p = 4$. The DPT allows any type of connectivity among the parts and our implementation applies a fully-connected constellation for maximally constrained geometry. The foreground/background models C_t are HSV color histograms with $16 \times 16 \times 16$ bins. The remaining parameters are as follows: the rate of spring system update is $\alpha_{\text{spr}} = 0.95$, the background histogram extraction area parameter is set to $\alpha_{\text{sur}} = 1.6$ and the histogram update rate is set to $\alpha_{\text{hist}} = 0.05$. These parameters have a straight-forward interpretation, were set to the values commonly used in published related trackers. Recall that the color informativeness test from Section 2.2.1 detects drastic segmentation failures. In our implementation the failure is detected if the number of pixels assigned to the object relative to the target bounding box size either falls below 20 percent or exceeds the initial size by 100 percent, i.e., $\alpha_{\min} = 0.2$ and $\alpha_{\max} = 2.0$. Note that these are very weak

constraints meant to detect obvious segmentation failures and did not require special tuning. The parameters have been fixed throughout all experiments.

The DPT was implemented in Matlab with backprojection and HoG extraction implemented in C and performed at 19 FPS on an Intel Core i7 machine. Since our tracker uses a KCF [11] for root and part appearance models, the complexity of our tracker is in order of the KCF complexity, which is $\mathcal{O}(n \log n)$, where n is the number of pixels in the search region. The DPT has complexity five times the KCF, because of the four mid-level parts plus a root part. The localization and update of five KCFs takes approximately 40ms. Our tracker consists also of the spring system and object segmentation. The optimization of the spring system takes on average less than 3ms and the color segmentation with the histogram extraction requires approximately 9ms.

3.2 The DPT design analysis

3.2.1 Analysis of the spring system optimization

This section analyzes the iterated direct approach (IDA) from Section 2.4, which is the core of our part-based optimization. The following random spring system was used in the experiments. Dynamic nodes were initialized at uniformly distributed positions in a 2D region $[0, 1] \times [0, 1]$. Each node was displaced by a randomly sampled vector $\mathbf{d} = [d_x, d_y] \sim \mathcal{U}([-0.5; 0.5])$ and the anchor nodes were set by displacing the corresponding dynamic nodes by the vector $\mathbf{b} = [b_x, b_y] \sim \mathcal{U}([-0.25; 0.25])$. The stiffness of i -th dynamic spring was set to $k_i = (\sigma d_i)^{-2}$, where d_i is the length of the spring and $\sigma = 0.1$ is the size change. The stiffness of j -th static spring was set to $k_j = \frac{1}{2} + u_j \bar{k}_{\text{dyn}}$, where \bar{k}_{dyn} is the average stiffness of the dynamic springs and $u_j \sim \mathcal{U}([0; 1])$. The IDA was compared with the widely used conjugate gradient descent optimization (CGD), which guarantees a global minimum will be reached on a convex cost function and has shown excellent performance in practice on non-convex functions as well [46]. All results here are obtained by averaging the performance on 100,000 randomly generated spring systems.

The first experiment evaluated the convergence properties of IDA. Figure 7 shows the energy reduction in spring system during optimization for different number of nodes in the spring system. The difference in the remaining energy after many iterations is negligible between CGD and IDA, which means that both converged to equivalent solutions. But the difference in energy reduction in consecutive steps and the difference in steps required to reach convergence is significant. The IDA reduces the energy at much faster rate than CGD and this result is consistent over various spring system sizes. Notice that IDA significantly reduced the energy already within the first few iterations.

The numeric behavior of IDA is much more robust than that of the CGD. Figure 8 shows an example of a spring system, where CGD did not reach the optimal state, but the IDA converged to a stable state with much lower energy, than the CGD. The poor convergence in CGD is caused by the very small distance between a pair of nodes compared to the other distances resulting in poor gradient estimation, while the IDA avoids this by the closed-form solutions for the marginal 1D spring systems. The IDA

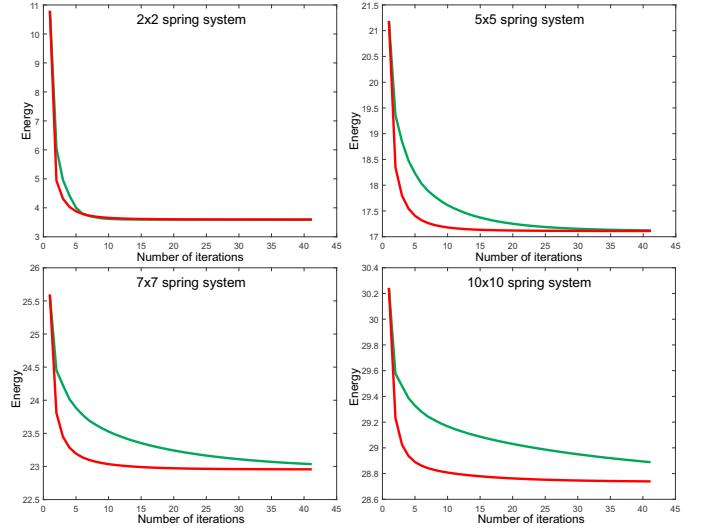


Fig. 7. The spring system remaining energy w.r.t. the iterations. Experiment is averaged over 10,000 random spring systems. The red and the green curves represent IDA and CGD methods, respectively.

converged in 5 iterations, while the CGD stopped after 471 iterations. The spring systems like the one described here were automatically detected and removed in the simulated experiment to prevent skewing results for the CGD. The results conclusively show that the IDA converges to a global faster than CGD and is more robust.

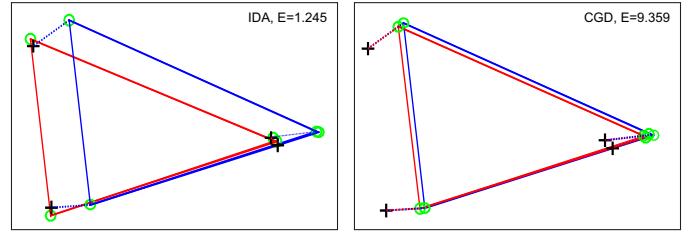


Fig. 8. The dynamic part of the spring system before and after optimization is shown in blue and red, respectively. Dynamic nodes and anchor nodes are depicted by green circles and black crosses, respectively, and the black dotted lines depict the static springs. The remaining energy E of the optimized spring system is shown as well.

The second experiment evaluated the IDA scalability. Figure 9 shows the optimization speed w.r.t. the spring system size. The number of iterations significantly increases for the CGD with increasing the number of parts. On the other hand, the IDA exhibits remarkable scalability by keeping the number of steps approximately constant over a range of system sizes. Furthermore, the variance in the number of iterations is kept low and consistently much lower than for the CGD. The iteration step complexity is expected to increase with the number of parts, since larger systems are solved. Figure 9 also shows that the computation times indeed increase exponentially for CGD, but the IDA hardly exhibits increase for a range of spring system sizes. These results conclusively show that IDA scales remarkably well.

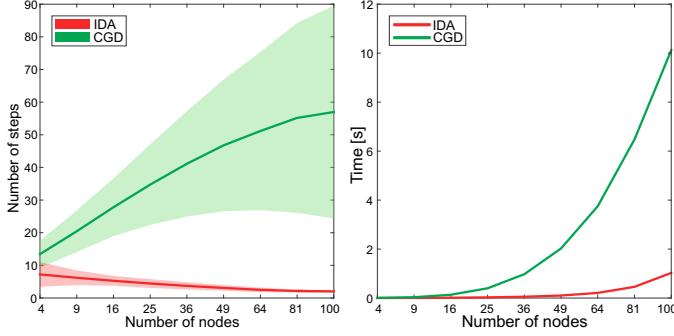


Fig. 9. The number of iterations (left) and time (right) spent by IDA and CGD on optimization with respect to the spring system size.

3.2.2 The DPT parameters analysis

The DPT design choices were evaluated on a state-of-the-art short-term tracking benchmark VOT2014 [4], [47]. In contrast to related benchmarks that aim at large datasets, the datasets in VOT initiative [47] are constructed by focusing on the challenging, well annotated, sequences while keeping the dataset small. The objects are annotated by rotated bounding boxes and all sequences are per-frame annotated by visual attributes. The VOT evaluation protocol initializes the tracker from a ground truth bounding box. Once the overlap between the ground truth and tracker output bounding box falls to zero, a failure is detected and tracker is re-initialized. The VOT toolkit measures two basic tracking performance aspects: reset-based accuracy and robustness. The reset-based accuracy is measured as the average overlap during successful tracking, while the robustness measures the number of failures (i.e., number of tracker re-sets). Apart from reporting raw accuracy/robustness values, the benchmark can rank trackers with respect to these measures separately by taking into account the statistical as well as practical difference. Since 2015 the VOT primary overall accuracy measure is the expected average overlap (EAO). This measure calculates the expected overlap on fixed-length sequences that a tracker would attain without reset. In addition we also report the primary OTB [1] measure. The OTB performance evaluation primarily differs from the VOT [41] in that trackers are not reset at failure. The overall performance is reported by an average overlap (AO) over all sequences.

The first experiment analyzed the contributions of the proposed segmentation in the coarse layer and the lower-layer constellation model. The baseline tracker was a DPT variant that does not use the constellation, nor the segmentation (DPT_{crs}^{nos}), which is in fact the original KCF [11] correlation filter. Adding a segmentation model to the baseline tracker results in the coarse layer in our part-based tracker, which we denote by DPT_{crs} . Table 1 clearly shows that the number of failures is reduced by our segmentation and the overall accuracy (EAO and AO) increases for DPT_{crs}^{nos} . By adding the lower layer to the DPT_{crs} , we arrive at the proposed DPT, which further boosts the performance by all measures. In particular, the number of failures is reduced by over 4%, the reset-based accuracy increases by over 10%, the expected average overlap (EAO) increases by 8% and the OTB average overlap (AO) increases by 10%. The VOT

ranking methodology was applied to these three trackers. The DPT was ranked as the top-performing tracker, which conclusively shows that the improvements are statistically as well as practically significant.

TABLE 1
Performance of DPT variants in terms of raw reset-based accuracy (res. acc.) and robustness (rob.), the VOT rank, the VOT no-reset accuracy (expected average overlap, EAO) and the OTB no-reset average overlap (AO). The arrows \uparrow and \downarrow indicate that “higher is better” and “lower is better”, respectively.

| DPT variant | VOT EAO \uparrow | Raw values | | VOT rank \downarrow | OTB AO \uparrow |
|--------------------------|--------------------|----------------------|-------------------|-----------------------|-------------------|
| | | res. acc. \uparrow | rob. \downarrow | | |
| DPT | 0.39 | 0.61 | 0.47 | 1.42 | 0.486 |
| DPT_{crs} | 0.36 | 0.55 | 0.49 | 2.10 | 0.442 |
| DPT_{crs}^{nos} | 0.21 | 0.57 | 1.13 | 3.06 | 0.377 |
| DPT_{str} | 0.34 | 0.57 | 0.61 | 2.06 | 0.467 |
| DPT_{loc} | 0.36 | 0.62 | 0.65 | 1.46 | 0.485 |
| $DPT_{ov}^{3 \times 3}$ | 0.31 | 0.60 | 0.71 | 1.82 | 0.481 |
| $DPT_{nov}^{3 \times 3}$ | 0.31 | 0.60 | 0.73 | 1.90 | 0.481 |

The DPT variants with fully connected, locally connected and star-based topology, DPT, DPT_{loc} , DPT_{str} , respectively, were compared to evaluate the influence of the lower-layer topology. The top performance in terms of the VOT EAO as well as OTB AO is achieved by the fully-connected topology, followed by the locally-connected and star-based topology. This order remains the same under the VOT ranking methodology, which confirms that the improvements of the fully-connected topology over the alternatives are statistically as well as practically significant.

For completeness, we have further tested the DPT performance with the increased number of parts at the lower layer. Given the constraints imposed on the parts size (as discussed in Section 3.1), we tested two variants with $3 \times 3 = 9$ parts: one with overlapping parts of the same size as in the original DPT ($DPT_{3 \times 3}^{ov}$) and one with smaller, non-overlapping, parts ($DPT_{3 \times 3}^{nov}$). Table 1 shows that these versions of DPT perform similarly in terms of overall performance (EAO and AO), with $DPT_{3 \times 3}$ obtaining slightly better rank, which is due to slightly better robustness than $DPT_{3 \times 3}^{ov}$. Both variants are outperformed by the original 2×2 DPT. The improvement of DPT over the best $DPT_{3 \times 3}$ tracker is over 20% in terms of the expected average overlap and approximately 2% in terms of the OTB average overlap. The smaller difference in OTB AO is because $DPT_{3 \times 3}$ has a similar accuracy as DPT, but fails more often. The OTB AO effectively measures the accuracy only up to the first failure. But the raw values clearly show superior robustness in DPT which is reflected in EAO.

3.3 Comparison to the state-of-the-art baselines

The DPT tracker is a layered deformable parts correlation filter, therefore we compared it to the state-of-the-art part-based as well as holistic discriminative trackers. The set of baselines included: (i) the recent state-of-the-art part-based baselines, PT [18], DGT [17], CMT [48] and LGT [13], (ii) the state-of-the-art discriminative baselines TGPR [26], Struck [5], DSST [9], KCF [11] SAMF [10], STC [32], MEEM [49], MUSTER [33] and HRP [50], and (iii) the standard baselines CT [23], IVT [22], MIL [6]. This

is a highly challenging set of recent state-of-the-art containing all published top-performing trackers on VOT2014, including the winner of the challenge DSST [9] and trackers recently published at major computer vision conferences and journals.

The AR-raw, AR-rank and the expected average overlap plot of the VOT2014 reset-based experiment are shown in Figure 10(a,b,c). In terms of AR-raw and AR-rank plots, the DPT outperforms all trackers by being closest to the top-right part of the plots. The tracker exhibits excellent tradeoff between robustness and accuracy, attaining high accuracy during successful tracks and rarely fails. This is reflected in the average expected overlap measure, which ranks this tracker as a top performing tracker (Figure 10c and the last row in Table 2). The DPT outperforms the best part-based tracker LGT [13] that applies a locally-connected constellation model and color segmentation by over 18% and the winner of the VOT2014 challenge, the scale adaptive correlation filter DSST [9], by 30%.

The VOT reset-based methodology resets the tracker after failure, but some trackers, like MUSTER [33], MEEM [49] and CMT [48] explicitly address target loss and implement mechanisms for target re-detection upon drifting. Although these are long-term capabilities and DPT is a short-term tracker that does not perform re-detection, we performed the no-reset OTB [1] experiment to gain further insights. The OTB [1] methodology reports the tracker overlap precision with respect to the intersection thresholds in a form a success plot (Figure 10d). The trackers are then ranked by the area under the curve (AUC) measure, which is equivalent to a no-reset average overlap [51]. The DPT outperforms the best baseline color-based superpixel short-term tracker DGT [17] and the long-term tracker MUSTER [33], which combines robust keypoint matching, correlation filter (DSST [9]), HoG and color features. The DPT also outperformed the recent state-of-the-art discriminative correlation filter-based trackers like DSST [9], color-based SAMF [10], the recently proposed multi-snapshot online SVM-based MEEM [49] and the recent logistic regression tracker HRP [50] tracker. The results conclusively show top global performance over the related state-of-the-art with respect to several performance measures and experimental setups.

3.3.1 Per-attribute analysis

Next we analyzed tracking performance with respect to the visual attributes. The VOT2014 benchmark provides a highly detailed per-frame annotation with the following attributes: *camera motion*, *illumination change*, *occlusion*, *size change* and *motion change*. In addition to these, we manually annotated sequences that contained deformable targets by the *deformation* attribute. If a frame did not contain any attribute or deforming target, it was annotated by an *empty* attribute.

The tracking performance with respect to each attribute is shown in Figure 11 and Table 2. The DPT outperforms all trackers on occlusion, camera motion, motion change and deformation and is among the top-performing trackers on illumination change, size change and empty. Note that the DPT outperformed all trackers that explicitly address target drift and partial occlusion, i.e., MUSTER [33], MEEM [49],

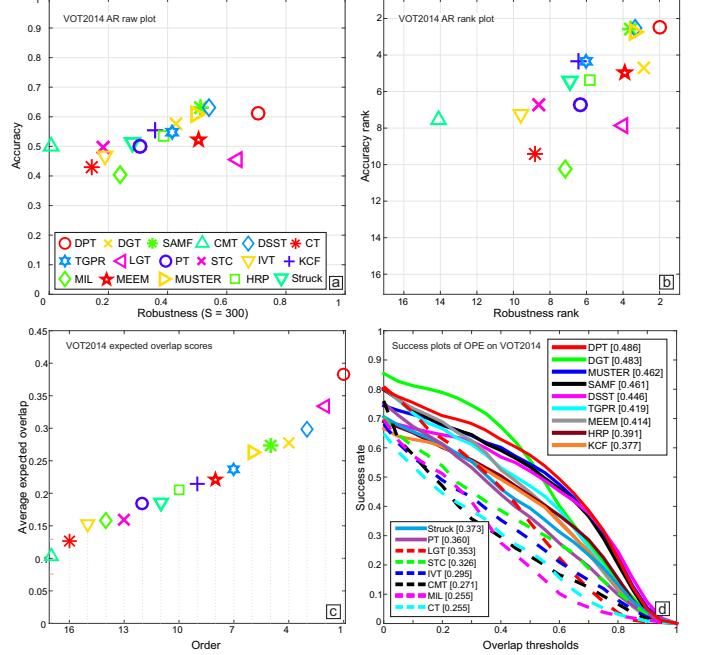


Fig. 10. The VOT2014 AR raw (a), AR rank (b), expected average overlap (c) and the OTB success plot (d).

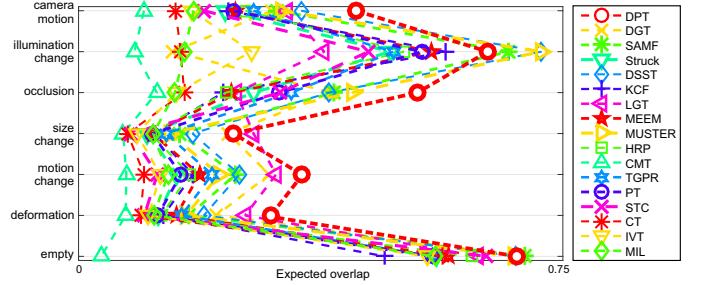


Fig. 11. The expected average overlap with respect to the visual attributes on the VOT2014 dataset.

CMT [48], Struck [5]. The DPT also outperforms top part-based trackers that address non-rigid deformations, i.e., LGT [13], DGT [17], PT [18] and CMT [48]. These results indicate a balanced performance in that the DPT does not only excel at a given attribute but performs well over all visual attributes.

3.4 Performance on benchmarks

For completeness of the analysis we have benchmarked the proposed tracker on the recent benchmarks. The DPT performance on the VOT2014 benchmark [4] compared to the 38 trackers available in that benchmark is shown in Figure 12. The DPT excels in the reset-based accuracy, robustness as well as the expected average overlap accuracy measure and is ranked third, outperforming 92% of the trackers on the benchmark. The two trackers that outperform the DPT are variants of the unpublished PLT tracker [4].

The DPT performance on the most recent and challenging VOT2015 benchmark [41] compared to the 60 trackers included in that benchmark are shown in Figure 13. The

TABLE 2

The per-attribute expected average overlap, i.e., EAO measure, (Ω), reset-based overlap (O) and number of failures (F) for the top 10 ranked trackers over 7 visual attributes: camera motion (CM), deformation (DE), empty (EM), illumination change (IC), motion change (MC), occlusion (OC), size change (SC). The arrows \uparrow and \downarrow indicate that “higher is better” and “lower is better”, respectively.

| attr. | DPT | | | LGT [13] | | | DSST [9] | | | DGT [17] | | | SAMF [10] | | | MUSTER [33] | | | TGPR [26] | | | MEEM [49] | | | KCF [11] | | | HRP [50] | | |
|---------|------|------|-------|-------------------|------|-------|-------------------|------|-------|-------------------|------|-------|-------------------|------|-------|-------------------|------|-------|-------------------|------|-------|-------------------|------|-------|-------------------|------|-------|----------|------|-------|
| | EAO↑ | DPT | F↓ | $\Omega \uparrow$ | O↑ | F↓ | | | |
| CM | 0.43 | 0.64 | 12.00 | 0.32 | 0.44 | 15.20 | 0.34 | 0.66 | 20.00 | 0.30 | 0.56 | 19.00 | 0.31 | 0.65 | 24.00 | 0.31 | 0.63 | 22.00 | 0.27 | 0.57 | 27.27 | 0.24 | 0.53 | 25.00 | 0.23 | 0.57 | 34.00 | 0.26 | 0.58 | 30.00 |
| DE | 0.31 | 0.60 | 17.00 | 0.26 | 0.41 | 11.16 | 0.19 | 0.56 | 28.00 | 0.21 | 0.54 | 16.00 | 0.17 | 0.59 | 31.00 | 0.17 | 0.55 | 31.00 | 0.16 | 0.53 | 37.07 | 0.15 | 0.52 | 33.00 | 0.13 | 0.53 | 42.00 | 0.13 | 0.50 | 41.00 |
| EM | 0.68 | 0.49 | 0.00 | 0.62 | 0.51 | 0.00 | 0.68 | 0.54 | 0.00 | 0.68 | 0.67 | 0.00 | 0.69 | 0.56 | 0.00 | 0.67 | 0.51 | 0.00 | 0.55 | 0.41 | 0.00 | 0.57 | 0.47 | 0.00 | 0.47 | 0.56 | 0.00 | 0.61 | 0.26 | 0.00 |
| IC | 0.64 | 0.63 | 1.00 | 0.38 | 0.45 | 1.47 | 0.72 | 0.74 | 1.00 | 0.15 | 0.46 | 14.00 | 0.67 | 1.00 | 0.72 | 0.73 | 1.00 | 0.49 | 0.57 | 3.47 | 0.55 | 0.54 | 2.00 | 0.57 | 0.54 | 1.00 | 0.50 | 0.66 | 4.00 | |
| MC | 0.35 | 0.63 | 14.00 | 0.31 | 0.46 | 10.47 | 0.25 | 0.64 | 24.00 | 0.30 | 0.58 | 14.00 | 0.24 | 0.66 | 25.00 | 0.22 | 0.64 | 26.00 | 0.21 | 0.55 | 30.20 | 0.19 | 0.53 | 24.00 | 0.18 | 0.57 | 34.00 | 0.17 | 0.60 | 35.00 |
| OC | 0.52 | 0.62 | 2.00 | 0.24 | 0.32 | 3.93 | 0.39 | 0.63 | 3.00 | 0.39 | 0.48 | 1.00 | 0.40 | 0.60 | 4.00 | 0.42 | 0.61 | 3.00 | 0.33 | 0.61 | 5.00 | 0.24 | 0.57 | 3.00 | 0.22 | 0.58 | 6.00 | 0.23 | 0.47 | 5.00 |
| SC | 0.24 | 0.54 | 12.00 | 0.27 | 0.43 | 7.40 | 0.18 | 0.52 | 15.00 | 0.23 | 0.57 | 6.00 | 0.16 | 0.56 | 18.00 | 0.14 | 0.53 | 19.00 | 0.14 | 0.47 | 21.20 | 0.11 | 0.46 | 15.00 | 0.12 | 0.47 | 27.00 | 0.11 | 0.50 | 27.00 |
| Average | 0.39 | 0.61 | 11.97 | 0.33 | 0.44 | 11.16 | 0.30 | 0.62 | 19.28 | 0.28 | 0.56 | 14.31 | 0.27 | 0.63 | 21.76 | 0.26 | 0.61 | 21.39 | 0.24 | 0.54 | 25.76 | 0.22 | 0.52 | 22.04 | 0.21 | 0.55 | 30.24 | 0.20 | 0.55 | 29.13 |

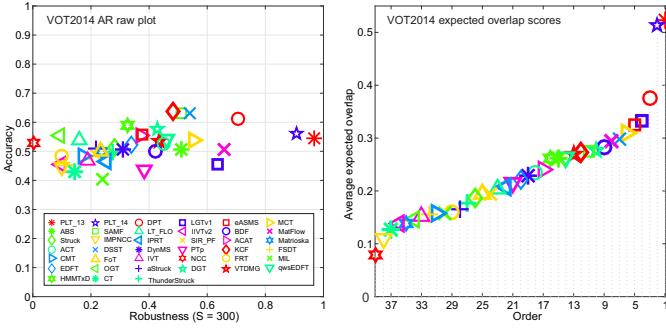


Fig. 12. The AR raw plots and the expected average overlap accuracy measures for VOT2014 benchmark [4]. Please see [4] for the tracker references.

tracker is ranked among the top 10% of all trackers, outperforming 54 trackers (i.e., 90% of the benchmark). The DPT outperforms all fifteen part-based trackers and fourteen correlation filter trackers, including the nSAMF, which is an improved version of [10] that applies color as well as fusion with various models, and the recently published improved Struck [52] that applies additional features and performs remarkably well compared to the original version [5]. The VOT2015 provides a *VOT2015 published sota bound* computed by averaging performance of trackers published in 2014/2015 in top computer vision conferences and journals. Any tracker with performance over this boundary is considered a state-of-the-art tracker according to VOT. The DPT is positioned well above this boundary and is considered a state-of-the-art according to the strict VOT2015 standards.

The DPT performance against 29 trackers available on the standard OTB [1] benchmark is shown in Figure 14. The DPT outperforms all trackers and is ranked top, exceeding the performance of the second-best tracker by over 8%.

3.5 Qualitative analysis

Qualitative analysis is provided for further insights. An experiment was performed to demonstrate the effectiveness of part adaptations during significant partial occlusions. The DPT was applied to a well-known sequence, in which the object (face) undergoes repetitive partial occlusions by a book (see Figure 15). The DPT tracked the face without failures. Figure 15 shows images of the face taken from the sequence along with the graph of color-coded part weights $w_t^{(i)}$. The automatically computed adaptation threshold is shown in gray. Recall that part is updated if the weight

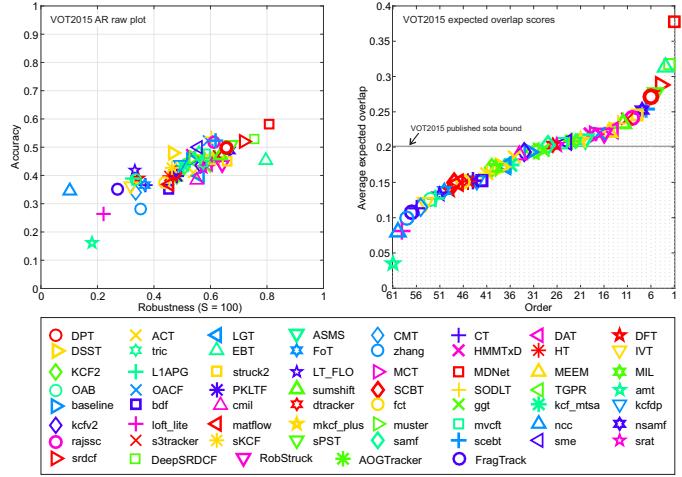


Fig. 13. The AR raw plots and the expected average overlap accuracy measures for VOT2015 benchmark [41]. Please see [41] for the tracker references.

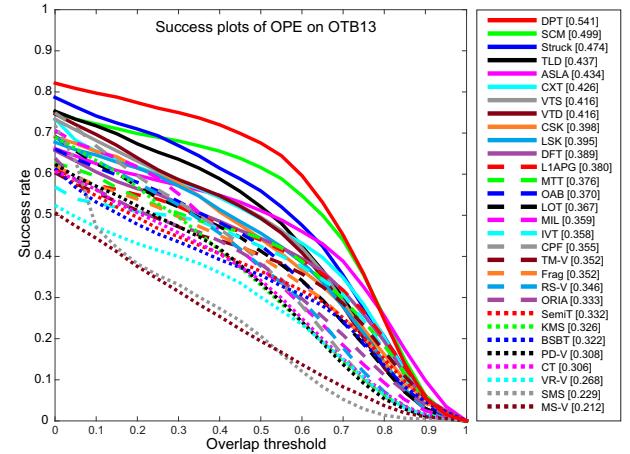


Fig. 14. The OPE performance plot for the top trackers on the OTB benchmark [1]. Please see [1] for the tracker references.

exceeds this threshold (Section 2.5). Observe that partial occlusions are clearly identified by the weight graphs, resulting in drift prevention and successful tracking through partial occlusions.

Additional qualitative examples are provided in Figure 16. The first row in Figure 16 shows performance on a non-deformable target with fast-varying local appearance.

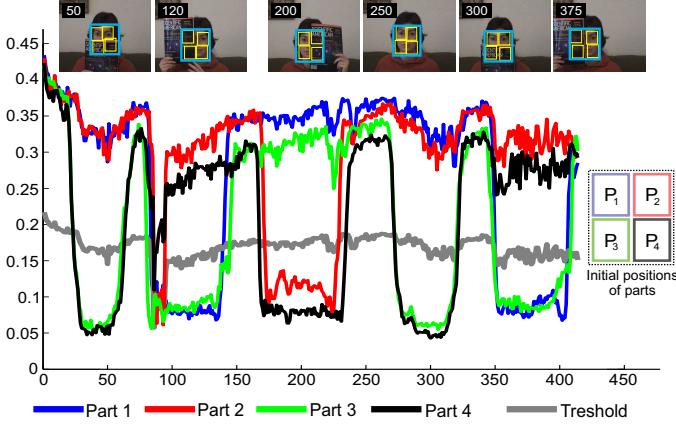


Fig. 15. Qualitative tracking results of partially occluded object. A sketch of parts is shown on the right-hand side. Part weights are color-coded, with the update threshold shown in gray.

The DPT tracks the target throughout the sequence, while holistic correlation- and SVM-based trackers [5], [9], [33] fail. The second, third and fourth row show tracking of deformable targets of various degrees of deformation. The fourth row shows tracking of a gymnast that drastically and rapidly changes the appearance. Note that the DPT comfortably tracks the target, while the related trackers fail. The first and second row in Figure 17 visualizes successful tracking performance on targets undergoing significant illumination changes. The third row shows tracking through several long-term partial occlusions. Again, the DPT successfully tracks the target even though the bottom part remains occluded for a large number of frames. The constellation model overcomes the occlusion and continues tracking during and after the occlusions.

4 CONCLUSION

A new class of deformable parts trackers based on correlation filters is presented. The developed deformable parts model jointly treats the visual and geometric properties within a single formulation, resulting in a convex optimization problem. The parts appearance models are updated by online regression to result in Gaussian-like likelihood functions and the geometric constraints are modeled as a fully-connected spring system. We have shown that the dual representation of such a deformable parts model is an extended spring system and that minimization of the corresponding energy function leads to a MAP inference on the deformable parts model. A highly efficient optimization called iterated direct approach (IDA) is derived for this dual formulation. A deformable parts correlation filter tracker (DPT) is proposed that combines a coarse object representation with a mid-level constellation of deformable parts model in top-down localization and bottom-up updates.

The extensive analysis of the new spring-system optimization method IDA showed remarkable convergence and robustness properties. In particular, the IDA converges much faster than the conjugated gradient descent, is numerically more robust and scales very well with increasing the number of parts in the spring system. Our tracker was rigorously compared against the state-of-the-art with

respect to several performance measures and experimental setups against sixteen state-of-the-art baselines. The DPT tracker outperforms the related state-of-the-art part-based trackers as well as state-of-the-art trackers that use a single appearance model, including the winner of the VOT2014 challenge and runs in real-time. Additional tests show that improvements come from the fully-connected constellation and the top-down/bottom-up combination of the coarse representation with the proposed deformable parts model. The DPT tracker was benchmarked on three recent highly challenging benchmarks against 38 trackers on VOT2014 [4] benchmark, 60 trackers on VOT2015 [41] benchmark and 29 trackers on the OTB [1] benchmark. The DPT attained a state-of-the-art performance on all benchmarks. Note that, since five KCFs [11] are used in DPT, the speed reduction is approximately five times compared to the baseline KCF. But the boost in performance is significant. The DPT reduces the failures compared to the baseline KCF by nearly 60%, the expected average overlap is increased by over 80% and the OTB average overlap is increased by approximately 30% while still attaining real-time performance.

The proposed deformable parts model is highly extendable. The dual formulation of the deformable constellation and the proposed optimizer are generally applicable as stand-alone solvers for deformable parts models. The appearance models on parts can be potentially replaced with other discriminative or generative models or augmented to obtain a constellation of parts based on various features like key-points and parts of different shapes. The part-based models like flocks of features [35], key-point-based [37], [48] and superpixel-based [17] typically use more parts than the tracker presented in this paper. Our analysis shows that the proposed optimization of the deformation model scales well with the number of parts, and could be potentially used in these trackers as a deformation model. Parts could also be replaced with scale-adaptive parts, which could further improve scale adaptation of the whole tracker. Alternatively, saliency regions could be used to improve localization. One way to introduce the saliency is at the coarse layer and another to apply it at the parts localization. Since the model is fully probabilistic, it can be readily integrated with probabilistic dynamic models. These will be the topics of our future work.

REFERENCES

- [1] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Comp. Vis. Patt. Recognition*, 2013, pp. 2411–2418.
- [2] A. Smeulders, D. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, July 2014.
- [3] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Čehovin, G. Nebehay, G. Fernandez, and T. e. a. Vojir, "The visual object tracking vot2013 challenge results," in *Vis. Obj. Track. Challenge VOT2013, In conjunction with ICCV2013*, Dec 2013, pp. 98–111.
- [4] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, T. Vojir, and G. et al. Fernandez, "The visual object tracking vot2014 challenge results," in *Proc. European Conf. Computer Vision*, 2014, pp. 191–217.
- [5] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," in *Int. Conf. Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2011, pp. 263–270.

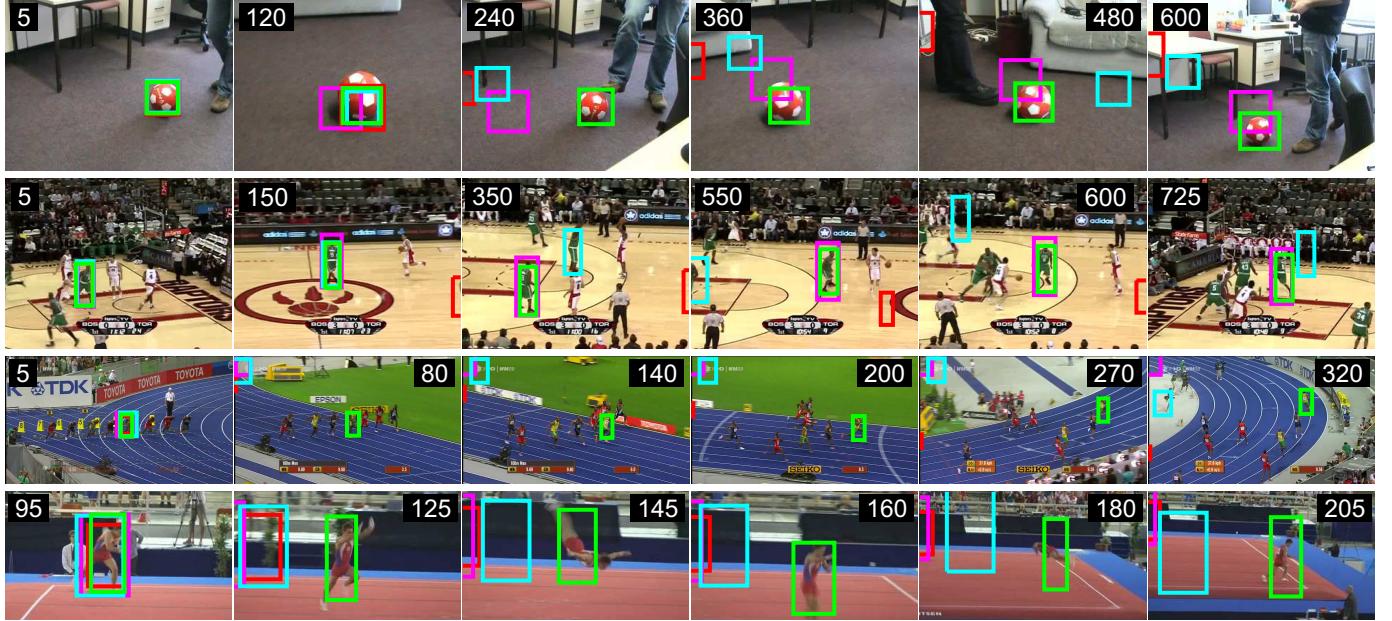


Fig. 16. Qualitative comparative examples of tracking for DPT, DSST, MUSTER and Struck shown in green, red, magenta and cyan, respectively.

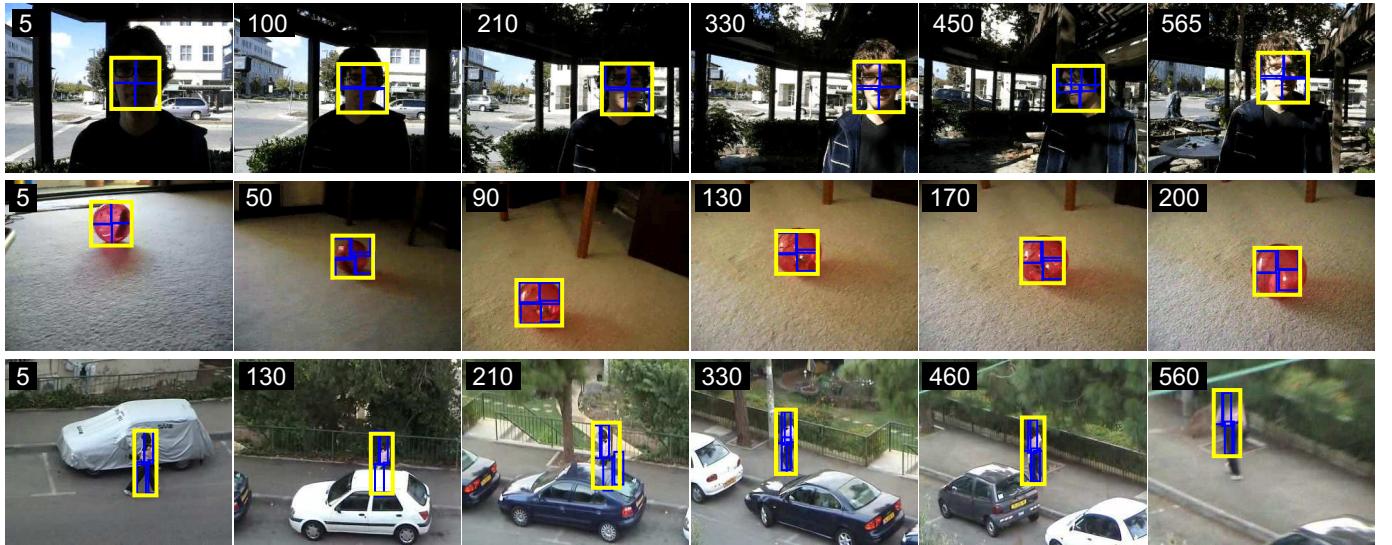


Fig. 17. Qualitative examples of DPT tracker on three sequences. Tracking bounding box is visualized with yellow color and four parts on mid-level representation are shown in blue.

- [6] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.
- [7] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proc. British Machine Vision Conference*, vol. 1, 2006, pp. 47–56.
- [8] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Comp. Vis. Patt. Recognition*. IEEE, 2010, pp. 2544–2550.
- [9] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. British Machine Vision Conference*, 2014, pp. 1–11.
- [10] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. European Conf. Computer Vision*, 2014, pp. 254–265.
- [11] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, 2014.
- [12] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.
- [13] L. Čehovin, M. Kristan, and A. Leonardis, "Robust visual tracking using an adaptive coupled-layer visual model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 941–953, Apr. 2013.
- [14] J. Kwon and K. M. Lee, "Tracking by sampling and integrating multiple trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1428–1441, July 2014.
- [15] M. E. Maresca and A. Petrosino, "Matrioska: A multi-level approach to fast tracking by learning," in *Proc. Int. Conf. Image Analysis and Processing*, 2013, pp. 419–428.

- [16] N. M. Artner, A. Ion, and W. G. Kropatsch, "Multi-scale 2d tracking of articulated objects using hierarchical spring systems," *Patt. Recogn.*, vol. 44, no. 4, pp. 800–810, 2011.
- [17] Z. Cai, L. Wen, Z. Lei, N. Vasconcelos, and S. Li, "Robust deformable and occluded object tracking with dynamic graph," *IEEE Trans. Image Proc.*, vol. 23, no. 12, pp. 5497 – 5509, 2014.
- [18] R. Yao, Q. Shi, C. Shen, Y. Zhang, and A. van den Hengel, "Part-based visual tracking with online latent structural learning," in *Comp. Vis. Patt. Recognition*, June 2013, pp. 2363–2370.
- [19] M. Godec, P. M. Roth, and H. Bischof, "Hough-based tracking of non-rigid objects," *Comp. Vis. Image Understanding*, vol. 117, no. 10, pp. 1245–1256, 2013.
- [20] G. Zhu, J. Wang, C. Zhao, and H. Lu, "Part context learning for visual tracking," in *Proc. British Machine Vision Conference*, 2014, pp. 1–12.
- [21] R. T. Collins, X. Liu, and M. Lordeanu, "Online selection of discriminative tracking features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1631–1643, 2005.
- [22] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vision*, vol. 77, no. 1-3, pp. 125–141, May 2008.
- [23] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking," in *Proc. European Conf. Computer Vision*, 2012, pp. 864–877.
- [24] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2259–2272, Nov 2011.
- [25] Z. Hong, X. Mei, D. Prokhorov, and D. Tao, "Tracking via robust multi-task multi-view joint sparse representation," in *Int. Conf. Computer Vision*, Dec 2013, pp. 649–656.
- [26] J. Gao, H. Ling, W. Hu, and J. Xing, "Transfer learning based visual tracking with gaussian processes regression," in *Proc. European Conf. Computer Vision*, vol. 8691, 2014, pp. 188–203.
- [27] S. Avidan, "Support vector tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 1064–1072, Aug 2004.
- [28] H. Possegger, T. Mauthner, and H. Bischof, "In defense of color-based model-free tracking," in *Comp. Vis. Patt. Recognition*, 2015, pp. 2113–2120.
- [29] P. Naidu, "Improved optical character recognition by matched filtering," *Optics Communications*, vol. 12, no. 3, pp. 287–289, 1974.
- [30] M. Zhang, J. Xing, J. Gao, and W. Hu, "Robust visual tracking using joint scale-spatial correlation filters," in *Proc. Int. Conf. Image Processing*. IEEE, 2015, pp. 1468–1472.
- [31] Y. Li, J. Zhu, and S. C. Hoi, "Reliable patch trackers: Robust visual tracking by exploiting reliable patches," in *Comp. Vis. Patt. Recognition*, 2015, pp. 353–361.
- [32] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Proc. European Conf. Computer Vision*. Springer International Publishing, 2014, pp. 127–141.
- [33] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking," in *Comp. Vis. Patt. Recognition*, 2015, pp. 749–758.
- [34] J. Hoey, "Tracking using flocks of features, with application to assisted handwashing," in *Proc. British Machine Vision Conference*, vol. 1, 2006, pp. 367–376.
- [35] T. Vojir and J. Matas, "The enhanced flock of trackers," in *Registration and Recognition in Images and Videos*, ser. Studies in Computational Intelligence. Springer Berlin Heidelberg, 2014, vol. 532, pp. 113–136.
- [36] B. Martinez and X. Binefa, "Piecewise affine kernel tracking for non-planar targets," *Patt. Recogn.*, vol. 41, no. 12, pp. 3682–3691, 2008.
- [37] F. Pernici and A. Del Bimbo, "Object tracking by oversampling local features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2538–2551, 2013.
- [38] X. Yang, Q. Xiao, S. Wang, and P. Liu, "Real-time tracking via deformable structure regression learning," in *Proc. Int. Conf. Pattern Recognition*, 2014, pp. 2179–2184.
- [39] S. Duffner and C. Garcia, "PixelTrack: a fast adaptive algorithm for tracking non-rigid objects," in *Int. Conf. Computer Vision*, 2013, pp. 2480–2487.
- [40] G. Duan, H. Ai, S. Cao, and S. Lao, "Group tracking: exploring mutual relations for multiple object tracking," in *Proc. European Conf. Computer Vision*. Springer, 2012, pp. 129–143.
- [41] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Čehovin, G. Fernandez, T. Vojir, G. Häger, G. Nebehay, and R. et al. Pflugfelder, "The visual object tracking vot2015 challenge results," in *Int. Conf. Computer Vision*, 2015.
- [42] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Comp. Vis. Patt. Recognition*, 2014, pp. 1090–1097.
- [43] M. Kristan, J. Perš, V. Sulic, and S. Kovačič, "A graphical model for rapid obstacle image-map estimation from unmanned surface vehicles," in *Proc. Asian Conf. Computer Vision*, 2014, pp. 391–406.
- [44] A. Diplaros, N. Vlassis, and T. Gevers, "A spatially constrained generative model and an em algorithm for image segmentation," *IEEE Trans. Neural Networks*, vol. 18, no. 3, pp. 798 – 808, 2007.
- [45] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Comp. Vis. Patt. Recognition*, vol. 1, June 2005, pp. 886–893.
- [46] I. Griva, S. G. Nash, and A. Sofer, *Linear and Nonlinear Optimization*, Second Edition. Siam, 2009.
- [47] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.
- [48] G. Nebehay and R. Pflugfelder, "Clustering of static-adaptive correspondences for deformable object tracking," in *Comp. Vis. Patt. Recognition*, 2015, pp. 2784–2791.
- [49] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: robust tracking via multiple experts using entropy minimization," in *Proc. European Conf. Computer Vision*, 2014, pp. 188–203.
- [50] N. Wang, J. Shi, D.-Y. Yeung, and J. Jia, "Understanding and diagnosing visual tracking systems," in *Int. Conf. Computer Vision*, 2015.
- [51] L. Čehovin, A. Leonardis, and M. Kristan, "Visual object tracking performance measures revisited," *IEEE Trans. Image Proc.*, vol. 25, no. 3, pp. 1261 – 1274, 2016.
- [52] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M. Cheng, S. Hicks, and P. Torr, "Struck: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.



Alan Lukežič received the Dipl.ing. and M.Sc. degrees at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia in 2012 and 2015, respectively. He is currently a researcher at the Visual Cognitive Systems Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Slovenia as a researcher. His research interests include computer vision, data mining and machine learning.



Luka Čehovin received his Ph.D from the Faculty of Computer and Information Science, University of Ljubljana, Slovenia in 2015. Currently he is working at the Visual Cognitive Systems Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Slovenia as a teaching assistant and a researcher. His research interests include computer vision, HCI, distributed intelligence and web-mobile technologies.



Matej Kristan received a Ph.D from the Faculty of Electrical Engineering, University of Ljubljana in 2008. He is an Assistant Professor at the Vi-CoS Laboratory at the Faculty of Computer and Information Science and at the Faculty of Electrical Engineering, University of Ljubljana. His research interests include probabilistic methods for computer vision with focus on visual tracking, dynamic models, online learning, object detection and vision for mobile robotics.