Discrete Texture Traces: Topological Representation of Geometric Context

Jan Ernst* and Maneesh K. Singh
Siemens Corporation, Corporate Research and Technology, Princeton, NJ, USA

Visvanathan Ramesh†
Department of Computer Science and Mathematics, Goethe University, Frankfurt am Main, Germany

Abstract

Modeling representations of image patches that are quasi-invariant to spatial deformations is an important problem in computer vision. In this paper, we propose a novel concept, the ***texture trace****, that allows sparse patch representations which are quasi-invariant to smooth deformations and robust against occlusions. We first propose a continuous domain model, the* ***profile trace****, which is a function only of the topological properties of an image and is by construction invariant to any homeomorphic transformation of the domain. We analyze its theoretical properties and then derive a discrete-domain approximation, the* ***Discrete Texture Trace*** *(****DTT****). DTTs are designed to be computationally practical and shown by a set of controlled experiments to be quasi-invariant to smooth spatial deformations as well as common image perturbations. We then show how DTTs can be naturally adapted to the incremental tracking problem, yielding highly precise results on par with the state of the art on challenging real data without using heavy machine learning tools. Indeed, we show that with even* just using one image *at the start of a sequence (i.e. no incremental updating), our method already outperforms four of six state of the art methods of the recent literature on challenging sequences.*

1. Introduction

Non-affine deformations are challenging image perturbations that arise in a variety of computer vision domains. They may be caused by viewpoint changes under perspective projection, or variations of objects that are deformable or articulated. The vision community has long sought representations that are quasi-invariant to such transformations for the use in detection, recognition and tracking. In contrast to affine spatial deformations, less focus has been directed at the weaker constraint of locally smooth deformations beyond the affine. Figure 2 shows an example for such deformations.

Figure 1 – One-shot tracking of the 'dudek' sequence [11]. The reference image is at the top left, the others show detections of our algorithm based on the reference. Visualizations are also available at `http://texturetraces.org`.

To introduce our concept, we note that many representations are explicitly or implicitly based on the notion of the Euclidean distance in the image domain. An image patch is often modeled as atomic image elements in particular spatial arrangements. Popular methods such as SIFT and shape context fall into this category as they essentially capture marginal distributions of edges in spatial configurations. For general deformations however, the Euclidean distance in the image domain may only be weakly related to the Euclidean distance in the original scene, e.g. on the surface of a curved object, especially over larger scales.

The main premise of this paper is that while Euclidean distances may change arbitrarily under smooth deformations, *the topology of the image is preserved*, i.e. the local neighborhood structure does not change. Instead of the Euclidean distance, we build our model based on *topological connectedness* and the preservation of local neighborhoods.

*Please direct correspondence to `jan.ernst@siemens.com`
†This work was performed while Visvanathan Ramesh was at Siemens.

After a literature review in section 2 we show in section 3 in a continuous space how this leads to a representation that is strictly invariant to smooth deformations of the image domain and provide a coarse discrete approximation that retains quasi-invariance in section 4. Then, in section 4.3, we represent an image patch as the set of all topological relations of its center location. Finally we show in section 5 how to use our concept for incremental tracking and point matching and compare to the state of the art in section 6. Section 7 concludes with final considerations and outlook.

2. Related Work

Feature descriptors that are invariant to geometric transformations have been addressed in a variety of ways. A popular stream of work is to estimate local parametric models of transforms, such as scale ([23, 20, 16]) and affinities ([24, 22, 12, 29]) based on the flat surface assumption. Several approaches treat feature detection and computation of the invariant feature descriptor as a single step ([6, 28]), or normalize with respect to a partial model and then capture the residual perturbations in the feature descriptor ([20, 5]). In [7] the authors combine multiple support regions into a classification scheme. Similarly, affine key point matching has been addressed by learning approaches ([3, 15]). The main difference of our approach to these methods is that we do not assume a parametric model of the projection or surface geometry, apart from smoothness of the deformation. Another important property of our approach is that we can compute a quasi-invariant descriptor at any textured point in the image, without requiring the detection of key points as a first step. Patch descriptions that are invariant to more general and non-parametric deformations have received much less attention. In [17] the authors pose the problem in a framework based on geodesic distances in the image. Our method makes much weaker assumptions on the underlying pattern. Both the recent chains model ([13]) and critical nets ([10]) are related in spirit in the sense that they use local contextual image information only to infer locations and encode invariant image features in a graph-based fashion. In [30] the authors present an invariant representation based on straight edge segments on surface discontinuities. We distinguish from landmark-based non-linear rectification ([31]) and earlier approaches that separately estimate the curvature of underlying surfaces ([21, 9, 2]) by not requiring the transformations to be parametric nor do we explicitly need to estimate such transformations.

Incremental tracking is an important problem and has been approached from a variety of directions (refer to the recent [18] for an overview) such as discriminative classifiers ([4]), generative models ([25]), combinations of trackers ([27]) and fragment representations ([1]). Our method has characteristics of fragment representations and also can offer generative understanding of local image regions.

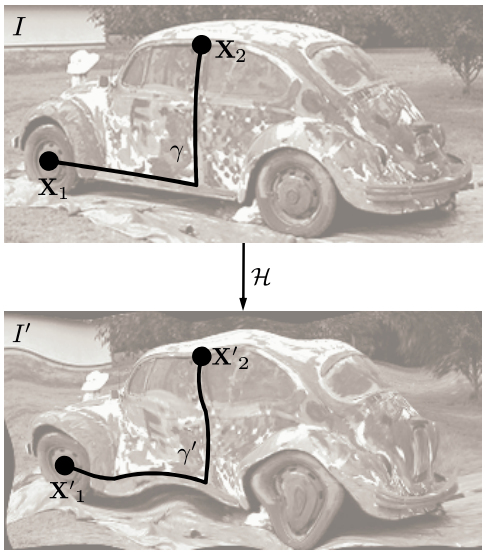


Figure 2 – An image I is spatially deformed by a homeomorphism $\mathcal{H}$. The points $\mathbf{X}_1$ and $\mathbf{X}_2$ are mapped into $\mathbf{X}'_1$ and $\mathbf{X}'_2$, and the spatial curve γ is mapped into γ'.

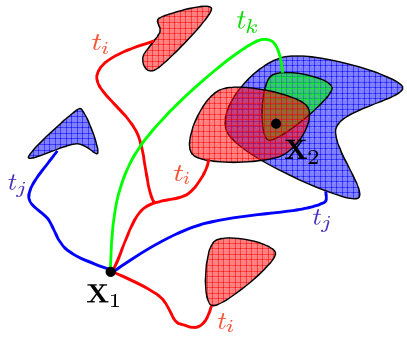


Figure 3 – Intersection of feasible sets for multiple traces $t_{i,j,k}$. Each trace induces a feasible set (coded by color) that can be reached given a start location $\mathbf{X}_1$. The location of $\mathbf{X}_2$ is constrained to lie in their intersection.

3. Continuous Profile Traces

3.1. Introduction

This section outlines the strategy followed in the rest of the paper. We motivate the approach in a continuous domain and then present a coarse discrete approximation in section 4. We begin by looking at two locations $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^2$ in images of a scene under a locally varying spatial deformation $\mathcal{H}$ as shown in figure 2. These deformations may stem from projective effects between a stereo image pair, non-rigid deformations of natural objects over time, intra-class variation of similar objects, etc.

Our approach is to model the location $\mathbf{X}_2$ by *describing how to get there* from $\mathbf{X}_1$ in a manner that depends only on the image texture and is invariant to local deformations.

Instead of defining $\mathbf{X}_2$ by a Euclidean (difference) vector from $\mathbf{X}_1$, we declare it in the same fashion one may use to follow directions in an unknown city: "To arrive at your destination from your current location, follow the road until you see a church on your right. Make a left at the next traffic light and drive until a roundabout. Take the second exit ..." and so on. This kind of guidance is of topological nature and (largely) invariant to smooth deformations of the underlying space and quantitative knowledge of the underlying metric is not necessary. In general this is not possible without ambiguities. This can be seen if one considers two locations in an image of constant gray level, where any spatial deformation will not present itself in the image texture. In effect, a particular description of how to get from $\mathbf{X}_1$ to $\mathbf{X}_2$ may also lead to other locations $\mathbf{X} \neq \mathbf{X}_2$. This partitions the image domain into two *equivalence sets*: the set of locations that can be reached by the description starting from $\mathbf{X}_1$ and the set of locations that can not be reached. We term the former the *feasible set* of $\mathbf{X}_1$ with respect to the particular description. This is visualized in figure 3. Different path descriptions t from $\mathbf{X}_1$ to $\mathbf{X}_2$ possibly induce different feasible sets. We can constrain the location of $\mathbf{X}_2$ the most by considering *all* feasible sets starting at $\mathbf{X}_1$ that contain $\mathbf{X}_2$ (cf. figure 3). If the path descriptions are invariant to local smooth perturbations, then by construction the feasible sets themselves as well as the constraint that the location of $\mathbf{X}_2$ is contained in their intersection are invariant.

The issue then is how to construct such an invariant path description? To this end we already noted that smooth deformations do not change the topology of the domain, i.e. the local neighborhood structure is preserved. This is precisely what we will use to construct our invariant description by assuming only the *preservation of local neighborhoods*, but not global spatial relations.

The following section 3.2 lays out the terminology and conceptual foundation in a continuous space.

3.2. Continuous Profile Traces

Although our concept is not limited to single channel images or even images at all, for clarity of exposure we make the following simplifying assumptions for the continuous construction: Pairs of continuous-space images $I, I' : \mathbb{R}^2 \to \mathbb{R}$ are spatially related as in figure 2 by a continuous mapping $\mathcal{H} : \mathbb{R}^2 \to \mathbb{R}^2$ with continuous inverse, i.e. a homeomorphism, and the brightness levels, contrast, etc. do not change between the images.

Now, in the domain of the image I, every curve $\gamma : [0, 1] \to \mathbb{R}^2$ with continuous velocity that connects two locations $\mathbf{X}_1$ and $\mathbf{X}_2$ (i.e. $\gamma(0) = \mathbf{X}_1$, $\gamma(1) = \mathbf{X}_2$) has an equivalent curve $\gamma'(r) = \mathcal{H}(\gamma(r))$ in the transformed image I' that connects the mapped points $\mathbf{X}'_1 = \mathcal{H}(\mathbf{X}_1)$ and $\mathbf{X}'_2 = \mathcal{H}(\mathbf{X}_2)$. The images I and I' as function of the curves γ and γ' respectively have the profiles $t(r) = I(\gamma(r)) = (I \circ \gamma)(r)$ and $t'(r) = (I' \circ \gamma')(r)$ which we call *profile traces* (or simply *traces* for short, as opposed to the curve γ). The traces have the property that $t(r) = t'(r)$ at every point r because $t'(r) = (I' \circ \gamma')(r) = (I' \circ \mathcal{H} \circ \gamma)(r) = (I \circ \gamma)(r) = t(r)$.

In other words the traces t do not change under smooth deformations. This property by itself is not very helpful as the deformation is exactly known in the construction of the trace t' and the curve γ' is not directly observable. A weaker but ultimately more useful statement is that *there exists some curve $\hat{\gamma}$ between $\mathbf{X}'_1$ and $\mathbf{X}'_2$ with the same trace* $\hat{t} = I' \circ \hat{\gamma} = t$. This is strictly a weaker criterion as the underlying curves $\hat{\gamma}$ and γ' are not necessarily the same. The existence property is not a function of $\mathcal{H}$ and thus invariant under $\mathcal{H}$.

Our goal is to restrict the true location of $\mathbf{X}'_2 = \mathcal{H}(\mathbf{X}_2)$ given the image I', a profile trace t and the location $\mathbf{X}'_1 = \mathcal{H}(\mathbf{X}_1)$. The following holds with regard to the location of $\mathbf{X}'_2$:

Proposition 1. *A necessary condition for any $\mathbf{X}'$ being the true location $\mathbf{X}' = \mathbf{X}'_2$ is the existence of a curve $\hat{\gamma}$ such that $\hat{\gamma}(0) = \mathbf{X}'_1$, $\hat{\gamma}(1) = \mathbf{X}'$ and the resulting trace $\hat{t}$ is equivalent to the trace t.*

As a semantic shortcut in the following we will write that the location $\mathbf{X}_2$ *is related to the location* $\mathbf{X}_1$ *by the trace* t (and: $\mathbf{X}'_2$ *is related to the location* $\mathbf{X}'_1$ *by the trace* t). Due to the potential ambiguities mentioned earlier, in general many locations $\mathbf{X}$ are related to $\mathbf{X}_1$ by a particular trace t. The set of these locations is what we call the *feasible set* of the trace t given the location $\mathbf{X}_1$. By construction, the induced feasible sets are invariant to $\mathcal{H}$ as each individual member of the set is defined in an invariant manner.

Finally, the best we can restrict the location of $\mathbf{X}'_2$ given $\mathbf{X}'_1$ is by the intersection of the feasible sets of *all traces* between $\mathbf{X}_1$ and $\mathbf{X}_2$ that start at $\mathbf{X}'_1$: If $\mathbf{X}_2$ is in the intersection of the feasible sets of all traces that lead from $\mathbf{X}_1$to $\mathbf{X}_2$, then $\mathbf{X}'_2$ is in the intersection of the feasible sets of the same traces starting from $\mathbf{X}'_1$.

4. Discrete Texture Traces

4.1. Discretization

In order to make the trace concept computationally practical, we need to make a set of approximations. The goal of the computation is to determine the feasible set given a start location and a trace description. Firstly we approximate the continuous curve γ as a finite sequence of discrete steps by quantizing the neighborhood around a location x_i as shown in figure 4a. There are n_θ angular bins and a neighborhood relation N. Two locations x_i and x_j are neighbors if they are closer than a neighborhood scale and they are then related by the particular discrete angular relation θ. In the continuous case, the local neighborhood

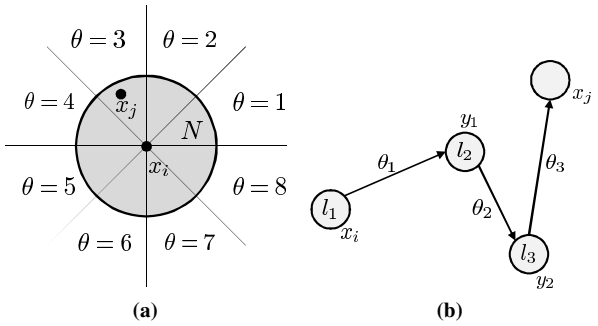


Figure 4 – (a) Neighborhood and angular relations: Location x_j is related to x_i by θ_3. (b) The discrete trace $\mathcal{T} = ((l_1, \theta_1), (l_2, \theta_2), (l_3, \theta_3))$ relates locations x_i and x_j

is considered relative to the orientation of the curve at that point. For computational purposes we look at two simpler alternatives: We either take the orientation relative to a common global coordinate system, yielding discrete traces that are not rotation invariant, or define the orientation relative to some estimate of the local texture orientation, yielding rotation invariant discrete traces. Finally, in order to represent the local image texture, we quantize the local image appearance into a set of discrete labels l. Any quantization yielding a robust labeling of local image appearance such as vector-quantized texture or shape descriptors would be suitable. Putting everything together, a discrete trace is a *finite sequence of discrete, labeled nodes x_i (with label l_i) and their discrete mutual neighborhood relations θ_i*. This very coarse discretization potentially introduces aliasing effects, implying that we do not retain the strict invariance of the continuous representation but quasi-invariance. The aliasing can be reduced by making the quantization finer.

4.2. Discrete Texture Traces

We are now equipped to formalize the discrete approximation of the continuous profile trace. Due to the fact that the labeled nodes capture local texture information, we term these *discrete texture traces* (DTT).

Definition 2. A discrete texture trace is a finite sequence of label-relation pairs $\mathcal{T} = ((l, \theta)_i)$ of length n_d. Given a starting location x_i it induces the feasible set of locations x_j that are reachable from x_i via the trace $\mathcal{T}$. A location x_j is reachable by $\mathcal{T}$ if there is a sequence of locations (y_k) such that $N(x_i, y_1) \wedge N(y_1, y_2) \wedge \ldots \wedge N(y_{n-1}, x_j) \wedge \theta(x_i, y_1) = \theta_1 \wedge \theta(y_1, y_2) = \theta_2 \wedge \ldots \wedge \theta(y_{n-1}, x_j) = \theta_n$ and the locations $(x_i, y_1, \ldots, y_{n-1})$ have labels $(l_1, l_2, \ldots, l_n)$ respectively.

Figure 4b illustrates this for the example DTT $\mathcal{T} = ((l_1, \theta_1), (l_2, \theta_2), (l_3, \theta_3))$ of length $n_d = 3$. The discrete neighborhood structure and labeled landmarks x, y induce a graph $\mathcal{G} = (E, V)$ with the relations N, θ as edges E and the landmarks as labeled nodes V. In order to construct the graph, the nodes y and x are sampled densely in the image domain. We define the set of *attributed adjacency matrices* $\{\mathbf{A}^{l\theta}\}$ of the graph $\mathcal{G}$ as: $a_{ij}^{l\theta} > 0$ iff the node i of label l has node j of arbitrary label connected to it by relation θ (as shown in figure 4a for $\theta = 3$). Then, according to definition 2 the trace $\mathcal{T} = ((l, \theta)_i)$ of length n_d relates the nodes x_i and x_j exactly if there is an intermediate sequence $(y_1, \ldots, y_{n-1})$ such that

$$a_{(x_i, y_1)}^{(l\theta)_1} \left(\prod_{k=2}^{n-1} a_{(y_{k-1}, y_k)}^{(l\theta)_k} \right) a_{(y_{n-1}, x_j)}^{(l\theta)_n} > 0 \qquad (1)$$

The matrices $\mathbf{A}$ are sparse and the existence of such a sequence can be established efficiently via matrix multiplication.

4.3. Patch Model

Now that we have shown how a specific DTT $\mathcal{T}$ relates two locations x_i and x_j we are ready to define how an image patch is generated: We simply model it by the *set of all texture traces* that have the patch's center point in their feasible set. In the continuous case of section 3 this is an infinite set and essentially describes the complete image domain in a topological manner. For the discrete approximation there is an implicit finite support region given by the discrete trace length n_d and the neighborhood scale of N. Formally, a patch centered at location x_j is modeled as the set $\{\mathcal{T}\}$ of all DTTs that have the location x_j in their feasible set for any other location x_i on the patch via equation 1.

In order to use DTTs for matching or detection we define the set $M = \{\mathcal{T}\}$ from one or more source image patches as the reference set or model. For a particular query location x_j in a target image we then compute the subset $\tilde{M} \subseteq M$ of traces that have x_j in their feasible set. Due to image perturbations that violate the modeling assumptions and the aforementioned aliasing, not all DTTs might be present in a transformed image. Accordingly we define the detection confidence of x_j as the relative number $|\tilde{M}| / |M|$ of reference traces whose existence for this location could be ascertained. Figure 6b shows the resulting confidence map, i.e. the intersection of the feasible sets sampled densely in the image domain, for the frame in figure 6a based on the reference set extracted from the top left patch in figure 1. The confidence map is generated by sampling the nodes x_j densely in the image and extracting the set $\tilde{M}$ for each of them. This confidence score between zero and one is used throughout the following experimental sections.

The computation of the feasible sets is highly parallelizable, because the individual traces can be computed independently.

Algorithm 1 Incremental tracking with discrete texture traces

1. Given an object's center location x_0 in the first frame, extract the set $R = \{\mathcal{T}\}$ of traces that lead to x_0. Set counter C for all possible DTTs $\mathcal{T}$: $C(\mathcal{T}) = 1$ if $\mathcal{T} \in R$, 0 otherwise.
2. Define $i = 0$, thresholds $t_d, t_r \in (0, 1)$, active set size n_a.
3. For new frame f
   (a) Predict $\tilde{x}_f$ linearly from x_{f-1}, x_{f-2}.
   (b) **Active set:** Define active set M^a as the n_a DTTs with the highest count $C(\mathcal{T})$. Ties are randomized.
   (c) **Detect:** Compute confidence map based on M^a at a dense set of locations in a window around $\tilde{x}_f$ and retrieve its maximum value c_{max} and location x_{max}.
   (d) **Detected:** If $c_{max} \geq t_d$, set $x_f = x_{max}$.
   (e) **Lost:** Else, store current model C as C^i, increment i, reset all $C_{\mathcal{T}} = 0$.
   (f) **Old models:** Compute confidence map for stored models $C^0 ... C^i$ individually, as in (b)-(c).
   (g) **Revert model:** If the confidence for any model C^j is higher than threshold t_r, set $C = C^j$ for the lowest such j, set $i = j$ and $x_f = x^j_{max}$.
   (h) **Update model:** Extract the set of traces that lead to the current center x_f and increase their count in C.

5. Incremental Tracking with Texture Traces

Incremental tracking of objects under perturbations such as illumination and scale changes, occlusion, in-plane and out-of-plane rotation and motion blur is an important, yet challenging problem. In this section, we show how our representation can be used to address some of these challenges in an incremental tracking framework. The tracking problem was chosen as an example primarily to demonstrate the representational power of the DTT. The basis for the algorithm is tracking by detection. In each new frame we compute the confidence for a dense set of locations in the image based on the current model M and use the location of the maximum confidence as the detected object position. Objects in video sequences may undergo gradual as well as rapid change of the their appearance. We address these changes by incrementally updating our model as well as dynamically keeping multiple models and possibly reverting to older models. The process is shown in algorithm 1.

6. Experiments and Results

Firstly we take a detailed look at synthetically perturbed data and examine the performance and invariance characteristics in comparison to geodesic intensity histograms (GIH, using the binaries provided by [17]) with respect to Gaussian noise, in-plane rotation, scale and smooth non-affine deformations. Secondly we tackle the challenging problem of incremental tracking under occlusion, in-plane and out-of-plane rotation and illumination change on realistic data.

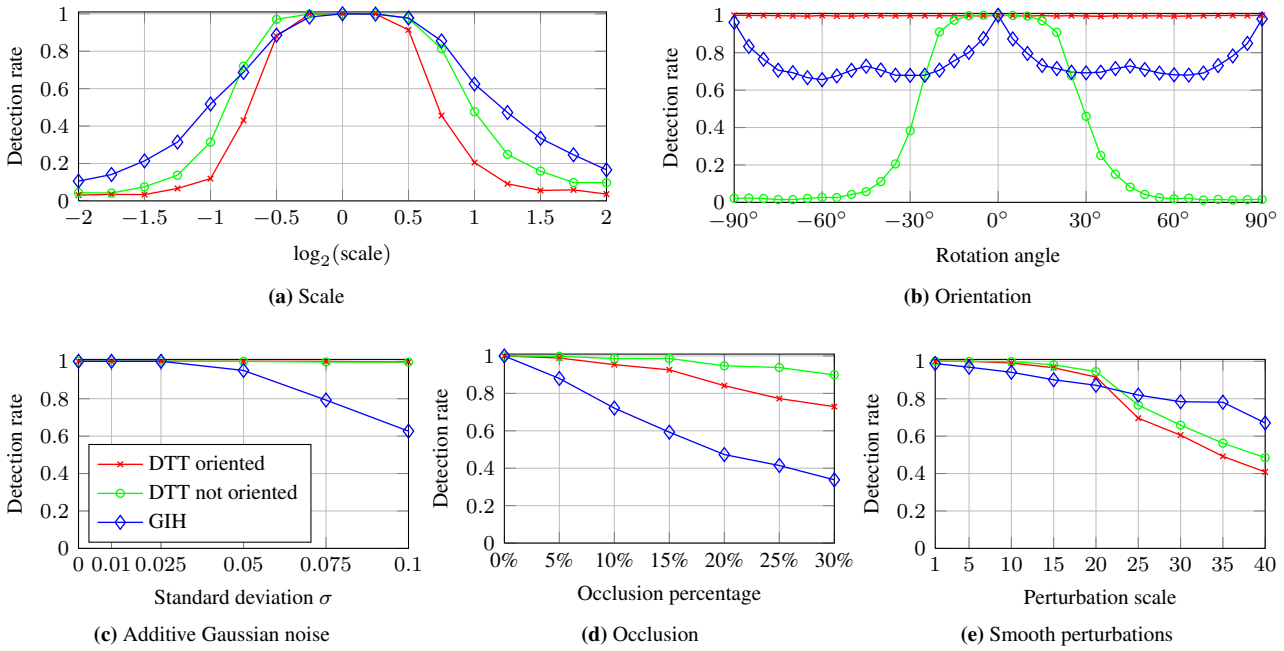
6.1. Base Performance and Occlusion Behavior

Performance Metric The detection rate we use as metric for this part has been used prior in [17]: For each pair of images, we select a set of key points in the first image and establish their corresponding ground truth locations in the second image. The detection rate is defined as the number of correct matches relative to the total number of key points in the reference image.

Experimental Setup All synthetic experiments were executed with the tuning parameters: trace length of $n_d = 3$, $n_\theta = 4$ possible angular quantization bins with a bin size of $\pm\pi/4$ and scale of neighborhood relation N of 20 pixels. These parameters were chosen empirically as a trade-off between performance and computation time. For smaller n_d the performance drops significantly, larger n_d become computationally challenging. We sample a dense set of intermediate node locations uniformly in the image and assign their labels by calculating SIFT descriptors at a fixed scale of 15 pixels square around each node and quantizing them into a fixed code book of size $n_l = 32$. For the oriented DTT, the orientation of the quantized patch is estimated as the major edge gradient orientation at the patch scale. The code book was determined by $k-$means from a set of descriptors gathered from a large image corpus unrelated to the test data. The GIH method was configured at $\alpha = 0.98$.

We randomly selected a set of 20 images from the PASCAL VOC 2010 database ([8]) and perturbed them to generate test images. We varied each of the five perturbation parameters Gaussian noise, in-plane rotation, scale, occlusion and smooth non-linear perturbation individually. For the stochastic parameters noise, occlusion and non-affine deformation we additionally generated three samples for each of the twenty images. The locally smooth perturbations were generated as multi-scale Perlin noise, varying the noise amplitude. Figure 2 illustrates a sample of the smooth perturbations. The occlusions were generated by randomly replacing 16 pixels square blocks of the test image by salt and pepper noise until a certain occlusion percentage was achieved. We found that the GIH performed significantly better if we smoothed the images with a Gaussian filter with a standard deviation of 0.75. As our method and the GIH do not require key point detection we sampled around 150 patch centers on a regular grid in the unperturbed image.

Discussion Figure 5 shows the performance as a function of the various perturbation parameters for our method and the GIH. The oriented and non-oriented DTT perform similar except in the case of rotation, where the rotational invariance of the oriented DTT comes into play. In the other cases, the oriented DTT performs slightly worse than the



(a) Scale **(b)** Orientation **(c)** Additive Gaussian noise **(d)** Occlusion **(e)** Smooth perturbations

Figure 5 – Performance on synthetic data as a function of the perturbation strength. The legend is shown in the lower left graph.

non-oriented DTT which may be explained by the trade off of the additional rotational invariance via possibly erroneous local orientation estimations. Both our methods clearly outperform the GIH for higher noise levels and occlusion. The noise performance may be explained by the smoothing of the underlying SIFT computation as well as the label quantization. The better occlusion behavior of the DTT may be explained by the voting-like confidence measure in the sense that an observed false-positive DTT does not negatively impact the result, but only non-observed reference DTTs impact the result by not voting. For small smooth deformations all methods perform similar, whereas the GIH outperforms both DTT versions for very large smooth deformations. This may be a result of the stricter invariance of the GIH to smooth spatial perturbations.

6.2. Incremental Tracking

We implemented the tracking algorithm described in section 5 and compared the performance of the non-rotation invariant DTT on eight challenging real-world sequences to several state of the art methods. We did not include the rotation invariant DTT as in-plane rotation is addressed by the incremental updating. Visualizations of the tracking performance are provided in the supplemental material and on-line at `http://texturetraces.org`. The empirically selected parametrization of the DTT was $n_\theta = 16$ possible angular bins with an orientation bin size of $\pm\pi/8$ and $n_l = 32$ discrete labels at a trace length of $n_d = 3$, resulting in $(n_\theta n_l)^{n_d} \approx 1.3 \times 10^8$ possible traces. We used a maximum number of about 0.3% of all possible traces as the active set size n_a, yielding a very sparse representation. The scale at which to track was automatically chosen as the scale at which the initial set M is the largest, i.e. the scale where the most unique traces can be computed for the initial patch. The eight sequences have a wide range of perturbations including motion blur, in- and out-of-plane rotation, occlusions and illumination change (Table 1 in [18] lists those individually). We used the two protocols from [18] and [27], and add to their comparisons respectively. The four sequences 'board', 'box', 'lemming' and 'liquor' of [27] were evaluated by the PASCAL score [8] against the current best performing SPT [18] as well as PROST [27], MIL [4], FragTrack [1], ORF[26] and GRAD [14]. We did not compare against the GIH as it is unclear how to extend this method to tracking. The PASCAL score measures the percentage of frames where the ground truth and detection overlap sufficiently to imply a correct detection. The results are shown in table 1. Our DTT method has a consistently high score and is on par with the SPT with an overall PASCAL performance of 95.5%. Furthermore we tested these four sequences and the four additional sequences 'david', 'girl', 'car' and 'faceocc2' that [18] use in their comparison. Their protocol uses the average deviation of the detected center from the ground truth relative to the diagonal size of the bounding box. The results are shown in table 2 in comparison to the same methods for the first four sequences and for the other four sequences in comparison to PROST, MIL, as well as TST [19] and IVT [25]. The DTT

Method	Average	board	box	lemming	liquor
PROST	80.4	75.0	90.6	70.5	85.4
MIL	49.2	67.9	24.5	83.6	20.6
FragTrack	66.0	67.9	61.4	54.9	79.9
ORF	27.3	10.0	28.3	17.2	53.6
GRAD	88.9	94.3	91.8	78.0	91.4
SPT	95.2	97.9	**94.8**	88.1	**100**
DTT	**95.5**	**99.3**	93.1	**91.4**	98.0
DTT one-shot	86.6	96.4	77.3	81.3	91.4

Table 1 – PASCAL score for the PROST sequences [27]. The best and second best method are highlighted respectively. We omit the distance score of [27] as a similar measure for these sequences is reported in table 2.

Method	lemming	box	board	liquor
PROST	0.189	0.091	0.157	0.101
MIL	0.112	0.740	0.206	0.619
FragTrack	0.625	0.406	0.363	0.145
ORF	1.256	1.030	0.623	0.304
GRAD	0.215	0.093	0.059	0.054
SPT	0.101	**0.073**	0.059	**0.016**
DTT	**0.094**	0.110	**0.050**	0.023

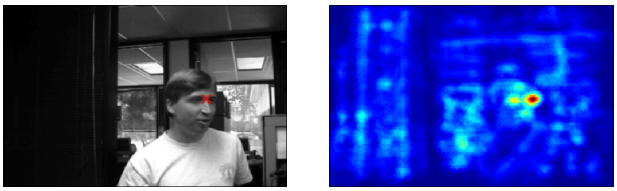
Method	david	girl	car	faceocc2
PROST	0.124	0.115	n/a	0.116
TST	0.052	0.131	0.065	0.139
MIL	0.127	0.161	0.700	0.095
IVT	0.059	0.147	**0.020**	0.081
SPT	**0.026**	**0.066**	0.031	0.065
DTT	0.033	0.105	0.022	**0.045**

Table 2 – Average tracking error relative to the diagonal size of the bounding box [18].

performs well compared to the recent SPT, particularly on the 'faceocc2' sequence which contains significant occlusion. The accuracy of our method on the 'box' sequence is limited by a lack of texture on the particular object. The run time of our method is on the order of tens of seconds per frame in Matlab.

6.3. One-shot Tracking

In order to get a better empirical understanding of the quasi-invariance properties of our representation, we asked the following question: how far can we get in tracking with using *only one frame for model building*, i.e. no incremental updating of the model. We call this *one-shot tracking* and it clearly stressed the invariance properties of any representation. We used the same tracking algorithm of section 5 but did not update the active set after the first frame. Furthermore we used all reference traces (i.e. $n_a = \infty$), computed the confidence map on the whole frame and suspended the



(a) Frame with detected maximum of confidence map.

(b) Confidence map

Figure 6 – One frame of the 'dudek' sequence [11] and the DTT confidence map based only on the first frame (shown in figure 1 top left, best viewed in color).

updating of x_f if $c_{max} < 0.1$ (i.e. less that 10% of the reference traces could be detected).

The resulting performance for the first four sequences is shown in table 1. When comparing the overall PASCAL performance of our one-shot method to the state of the art one can see that it outperforms four of the six compared methods. In other words, with just using one initial frame, our representation already takes third place out of seven, outperformed only by the recent GRAD and SPT methods. To illustrate the performance of our one-shot tracking, we applied it to the 'dudek' sequence [11]. Figure 1 shows at the top left the initial image cropped around the ground truth location and several detections of our algorithm throughout the sequence. Please note how the detected center point is always on the bridge of the nose between the eyes (as is the ground truth). Figure 6 shows one frame and the computed confidence map. The overall PASCAL performance on this sequence with our one shot tracking is 99.5%. The average distance between ground truth and detection is 2.5 pixels (or a ratio of 0.011 when measured as in table 2), indicating that our detections are very precise. We omit the one-shot results for table 2 since outliers skew the average ratio significantly, rendering this particular comparison meaningless.

7. Summary and Conclusions

We have described a new approach for representing deformable domains such as image patches, the *profile trace* and a particular discrete approximation, the *discrete texture trace* (DTT). We validated the DTT under perturbations including scaling, rotation, spatial deformation, occlusion and Gaussian noise. Furthermore we demonstrated a highly precise incremental tracking system based on our representation that is on par with the state of the art. Indeed, even without incremental updating, the DTT already outperforms four of six trackers of the recent literature. All of this is achieved without heavy machine learning tools or sophisticated tracking modules. The DTT representation is highly parallelizable, lending itself to GPU implementations.

Apart from using the DTT as a basis representation in other domains such as object class detection, we see two

immediate extensions of our method. Firstly, the attributed adjacency matrices are not restricted to only encoding spatial and quantized appearance relations. Additional domain specific information such as motion in tracking or segmentations in object detection are straightforward to incorporate. As an example, the prior assumption of common foreground motion in tracking can be encoded in our model by including a particular relation in the adjacency matrices only if the neighboring locations have similar motion. This implies that a particular trace cannot cross a boundary with larger motion difference. This leads to *motion consistent texture traces*. Secondly, a trace relates the center location to another point on the patch. The actual observed configuration of the trace can then aid in reasoning about the scene geometry. For instance, the *non-existence* of a particular trace implies that any or all intermediate or end nodes are not observed.

Finally, the trace concept is applicable beyond just images to a larger set of multidimensional signals.

Acknowledgements We would like to thank Yakup Genc of Siemens Corporate Research for valuable discussions.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. *CVPR*, 1:798–805, 2006. 2, 6

[2] J. Aloimonos. Shape from texture. *Biol. Cybern.*, 58(5):345–360, 1988. 2

[3] B. Babenko, P. Dollar, and S. Belongie. Task specific local region matching. In *ICCV*, pages 1–8, 2007. 2

[4] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *CVPR*, pages 983–990, 2009. 2, 6

[5] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.*, 110:346–359, June 2008. 2

[6] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE PAMI*, 24(4):509–522, Apr 2002. 2

[7] H. Cheng, Z. Liu, N. Zheng, and J. Yang. A deformable local image descriptor. In *CVPR*, pages 1–8, 2008. 2

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge. *IJCV*, 88(2):303–338, 2010. 5, 6

[9] J. Gårding and T. Lindeberg. Direct computation of shape cues using scale-adapted spatial derivative operators. *IJCV*, 17(2):163–191, 1996. 2

[10] S. Gu, Y. Zheng, and C. Tomasi. Critical nets and beta-stable features for image matching. In *ECCV*, pages 663–676. Springer-Verlag, 2010. 2

[11] A. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. *IEEE PAMI*, 25(10):1296–1311, 2003. 1, 7

[12] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *ECCV*, volume 3021, pages 228–241. Springer, 2004. 2

[13] L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The chains model for detecting parts by their context. In *CVPR*, pages 25 –32, 2010. 2

[14] D. Klein and A. Cremers. Boosting scalable gradient features for adaptive real-time tracking. In *ICRA*, pages 4411 –4416, May 2011. 6

[15] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *PAMI*, 28(9):1465–1479, 2006. 2

[16] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, 1998. 2

[17] H. Ling and D. W. Jacobs. Deformation invariant image matching. In *ICCV*, pages 1466–1473, 2005. 2, 5

[18] B. Liu, J. Huang, L. Yang, and C. Kulikowski. Robust tracking using local sparse appearance model and k-selection. In *CVPR*, pages 1313–1320, 2011. 2, 6, 7

[19] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, and C. Kulikowski. Robust and fast collaborative tracking with two stage sparse optimization. In *ECCV*, pages 624–637. Springer, 2010. 6

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[21] J. Malik and R. Rosenholtz. Computing local surface orientation and shape from texture for curved surfaces. *IJCV*, 23(2):149–168, 1997. 2

[22] J. Matas, O. Chum, U. Martin, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages 384–393, 2002. 2

[23] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV 2001*, volume 1, pages 525–531, 2001. 2

[24] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004. 2

[25] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *IJCV*, 77:125–141, May 2008. 2, 6

[26] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof. On-line random forests. In *ICCV*, pages 1393–1400, 2009. 6

[27] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof. PROST Parallel Robust Online Simple Tracking. In *CVPR*, 2010. 2, 6, 7

[28] D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *ECCV*, pages 814–828. Springer, 2000. 2

[29] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *Int. J. Comput. Vision*, 59(1):61–85, 2004. 2

[30] A. Vedaldi and S. Soatto. Features for recognition: Viewpoint invariance for non-planar scenes. In *ICCV 2005*, volume 2, pages 1474–1481, 2005. 2

[31] B. Zitová and J. Flusser. Image registration methods: a survey. *IVC*, 21(11):977–1000, 2003. 2