Locally Orderless Tracking

Shaul Oron[1] shauloro@post.tau.ac.il
Aharon Bar-Hillel[2] aharon.barhillel@gm.com
Dan Levi[2] dan.levi@gm.com
Shai Avidan[1] avidan@eng.tau.ac.il

(1) Tel Aviv University, Tel Aviv 69978, Israel
(2) General Motors Advanced Technical Center, Hamada 7, Herzliya, Israel

Abstract

Locally Orderless Tracking (LOT) is a visual tracking algorithm that automatically estimates the amount of local (dis)order in the object. This lets the tracker specialize in both rigid and deformable objects on-line and with no prior assumptions. We provide a probabilistic model of the object variations over time. The model is implemented using the Earth Mover's Distance (EMD) with two parameters that control the cost of moving pixels and changing their color. We adjust these costs on-line during tracking to account for the amount of local (dis)order in the object. We show LOT's tracking capabilities on challenging video sequences, both commonly used and new, demonstrating performance comparable to state-of-the-art methods.

1. Introduction

In visual tracking one often has to make an explicit or implicit assumption about the type of the object being tracked, treating it as either a rigid object or a deformable one. For example, if tracking a rigid object, where the only appearance change is due to rigid geometric transformations, it is reasonable to use a method such as template matching where pixel locations are fixed and governed by a geometric transformation and similarity is reduced to per-pixel intensity difference. If, on the other hand, the object is extremely deformable, then tracking based on color histogram matching might be more suitable reducing the similarity between target and candidate to similarity between their color distributions.

In this work we present Locally Orderless Tracking (LOT), a novel visual tracking algorithm that uses a joint spatial-appearance space and is able to estimate, on-line, the amount of local (dis)order in the target. Thus if the target is rigid and there is little or no local disorder then LOT preserves spatial information like template matching. However, if the target is nonrigid, LOT disregards spatial information as in histogram matching.

The first contribution of our work is a new probabilistic interpretation of the Earth Mover's Distance (EMD) that we name Locally Orderless Matching (LOM). Using LOM one can calculate the likelihood of patch P being a noisy replica of patch Q where noise can be introduced by change in the spatial order of pixels in the patch, change in their appearance, or both. In other words, LOM infers the probability $Pr(P|Q,\Theta)$ where Θ are noise model parameters, some of which control the cost of moving pixels spatially while others control the cost of changing a pixels appearance, for example due to illumination variation.

The second contribution of our work is introducing Locally Orderless Tracking which applies Locally Orderless Matching to visual tracking. This is done, in a generative approach, using particle filtering where particle likelihoods are inferred using LOM. Particles are represented as signatures in a joint spatial-appearance space, using superpixels for better efficiency. Key to our approach is the noise model used in LOM to regulate the cost of moving superpixels and changing their appearance, and we show how the optimal parameters of this noise model can be estimated on-line, using maximum likelihood optimization, according to the degree of "rigidity" in the object.

2. Related Work

We are inspired by the work of Koenderink and Van Doorn on the structure of locally orderless images [1] which proposes an image representation method where the amount of spatial order preserved globally and locally can be tuned using two parameters. This representation was shown by Ginneken and Haar Romeny [2] to be useful for applications such as adaptive histogram equalization, noise removal and segmentation. In our case, we wish to determine the optimal extent of local disorder of the data for the purpose of tracking.

In rigid object tracking one usually attempts to exploit spatial information in the object by using template based methods. In some cases the template is used in a simple manner [3] while others use multiple templates and sparse

representations [4, 5, 6, 7]. These approaches offer good stability and can handle occlusions and scale estimation but are less suitable for handling non-rigid deformations and dynamics such as out-of-plane-rotations.

When tracking deformable objects one often uses histogram representations [8] or discriminative methods that treat the problem as a pixel-wise binary classification problem [9, 10, 11]. These approaches mostly disregard spatial order, and can therefore handle difficult non-rigid transformations. However they are more prone to drift and are often less stable especially at scale estimation or occlusion handling. Some attempt to combine rigid and deformable object approaches either by using mid level cues that capture spatial information to some extent [12] or by heuristically combining discriminative and generative components [13]. However these methods do not measure the extent of local disorder in the data explicitly and adapt accordingly like LOT.

The work most related to ours is that of Elgammal *et al.* [14] that propose a tracker that uses a joint spatial-appearance space and can specialize to either histogram tracking or sum-of-square-difference (SSD) tracking by an off-line adjustment of parameters. The proposed method is significantly different in several ways. First and foremost, due to the on-line parameter estimation which enables LOT to specialize in rigid template tracking or deformable object tracking on-line and secondly due to the use of particle filtering and EMD instead of the kernel based gradient decent approach of Elgammal *et al.*

The Earth Mover's Distance (EMD) has a long history in computer vision. EMD was first considered by Peleg *et al.* [15] as an image similarity metric and popularized by Rubner *et al.* [16] (who coined the name) for content based image retrieval. A probabilistic analysis of EMD and its relation with the Mallows distance was proposed by [17] although that analysis differs from the proposed probabilistic framework which introduces a noise process that governs the ground distance in the EMD. Recently, Zhao *et al.* [18] proposed a differential EMD approach that derives a gradient descent method to find the object location quickly using the EMD as a similarity measure. However, the focus of that paper is on using EMD to handle illumination changes, the object is represented as a color signature and no consideration is given to pixel location.

Superpixels[19] have been used in recent years for many computer vision applications such as segmentation, classification [20, 21] and tracking [12]. In our work, similar to [22], superpixels are used to reduce the computational cost of EMD.

We refer interested readers to a thorough survey of the vast work in visual tracking that can be found in [23].

3. Locally Orderless Matching

Locally Orderless Matching measures the similarity between two images or two image patches based on the EMD. Pixels are represented in a joint spatial-appearance domain. For appearance we use color values but other descriptors such as local gradients or texture can also be used. For position pixel coordinates in a patch, normalized to the range $[0, 1]$, are taken. A pixel is represented as $p_i = (p_i^L, p_i^A)$ where $p_i^L = (x, y)$ is the pixels location and $p_i^A \in \mathbb{R}^k$ its appearance.

We want to probabilistically explain a candidate patch P as a noisy replica of the template Q. We begin by looking at the pixel-wise inference problem, where patches P and Q are treated as sets of pixels, and show that in this case the problem is equivalent to a form of EMD optimization problem. We then propose using signature representations for P and Q, in which superpixels are used to cluster pixels together, and claim the problem can now be formulated as the signature EMD problem [16]. This is done in order to reduce the computational cost of EMD and we justify by bounding the error resulting from the related coarsening of the representation.

Let us consider patches P and Q as sets of pixels. We start with a probabilistic perspective of EMD and wish to show that it measures the conditional probability of one set, given the other set and model parameters. Formally, denote the two sets by $P = \{p_i\}_{i=1}^n, Q = \{q_i\}_{i=1}^n$, and assume that we have a probabilistic model stating the probability that a specific element $p \in P$ originated from a specific element $q \in Q$, $Pr(p|q, \Theta)$, with Θ the model parameters. We want to extend it to the conditional probability between the sets $Pr(P|Q, \Theta)$.

The extension relies on a hidden 1:1 mapping between elements of P and Q. Denote such a mapping by $h : \{1, .., n\} \rightarrow \{1, .., n\}$ with $h(i) = j$ meaning that element p_i was generated from element q_j. We can get the probability of P being generated from Q by marginalizing over the possible hidden assignments (dropping Θ from the notation as it is currently constant):

$$Pr(P|Q) = \sum_h Pr(P|Q, h)Pr(h) \tag{1}$$

Assuming a uniform prior over the h's (no reason to assume anything else) we have:

$$Pr(P|Q) = \frac{1}{n!} \sum_h Pr(P|Q, h) \tag{2}$$

Approximating the average using maximum a posteriori (MAP) estimation, i.e. assuming the sum is dominated by the highest term (the best hidden map) we get:

$$Pr(P|Q) \sim c \cdot \max_h Pr(P|Q, h) \tag{3}$$

Dropping the constant c, assuming independence between the set elements and taking the logarithm we get:

$$\begin{array}{rcl} logPr(P|Q) & \sim & \max_h logPr(P|Q,h) \\ & = & \max_h \sum_{i=1}^{n} logPr(p_i|q_{h(i)},\Theta) \end{array} \quad (4)$$

Proposition 3.1 *Optimization problem* (4) *is the signature EMD problem EMD(P,Q,d) for the following signatures and ground distance:*

$$\begin{array}{rcl} P & = & \{(p_1,1),(p_2,1),\ldots,(p_n,1)\} \\ Q & = & \{(q_1,1),(q_2,1),\ldots,(q_n,1)\} \\ d(p,q) & = & -logPr(p|q,\Theta) \end{array} \quad (5)$$

Where the signatures are comprised of objects, e.g. (p_i, w_i)*, each having a description* p_i *and weight* w_i*. In our case the signatures are simply collections of all the pixels in patches* P *and* Q *equally weighted.*

Proof Starting with Equation (4) we have:

$$\begin{array}{l} \max_h \sum_{i=1}^{n} logPr(p_i|q_{h(i)},\Theta) \quad = \\ \min_h \sum_{i=1}^{n} -logPr(p_i|q_{h(i)},\Theta) \quad = \\ \min_h \sum_{i=1}^{n} d(p_i,q_{h(i)}) \end{array} \quad (6)$$

where the mapping h can be expressed as a permutation matrix F in which $f_{ij} = 1$ iff $h(i) = j$. Denoting $d_{ij} = d(p_i, q_j)$ the problem statement becomes:

$$\begin{array}{ll} & \min \sum_{i,j} f_{ij} d_{ij} \\ \text{such that} & \\ & \sum_i f_{ij} = 1, \sum_j f_{ij} = 1, f_{ij} \in \{0,1\} \end{array} \quad (7)$$

If we put this integer linear programming problem in the canonical form $\{\min c \cdot x | Ax = b, x \geq 0\}$ we find that the matrix A is totally unimodular [24]. This in turn implies that the linear programming problem in which we relax the constraint $f_{ij} \in \{0,1\}$ to $f_{ij} \geq 0$ has an integral optimum, meaning the constraint can be relaxed without changing the result.

The linear programming problem obtained by this relaxation is identical to the signature EMD with identical mass presented by Rubner *et al*. [16].

$$\begin{array}{ll} & \min \sum_{i,j} f_{ij} d_{ij} \\ \text{such that} & \\ & f_{ij} \geq 0, \sum_i f_{ij} \leq w_{q_j}, \sum_j f_{ij} \leq w_{p_i} \\ & \sum_{i,j} f_{ij} = \min(\sum_i w_{p_i}, \sum_j w_{q_j}) \end{array} \quad (8)$$

To see this notice that for signatures with identical mass the inequalities $\sum_i f_{ij} \leq w_{q_j}, \sum_j f_{ij} \leq w_{p_i}$ can be replaced by equalities and then the last constraint $\sum_{i,j} f_{ij} = \min(\sum_i w_{p_i}, \sum_j w_{q_j})$ can be dropped. ∎

In other words, conditional set probability, under 1:1 mapping and element independence assumptions, is equivalent to signature EMD with singleton bins. However, the equivalence naturally extends to conditional probabilities with P and Q containing repeating elements and signature EMD with general integer bin quantities.

Proposition 3.2 *Let*

$$\begin{array}{rcl} P & = & \{(p_1,w_1^p),(p_2,w_2^p),\ldots,(p_{n_1},w_{n_1}^p)\} \\ Q & = & \{(q_1,w_1^q),(q_2,w_2^q),\ldots,(q_{n_2},w_{n_2}^q)\} \end{array} \quad (9)$$

be signatures for which we cluster repeating elements into single objects increasing their weights accordingly (e.g. p_1 *appears* $w_1^p \in \mathbb{N}$ *times in P, etc.). Solving the 1:1 pixel matching problem for* P *and* Q *as formulated in equation* (7) *(which has* m^2 *variables where* $m = \sum_{i=1}^{n} w_i^p$*) is equivalent to solving the EMD problem* (8) *for* P *and* Q *(which has* $n_1 \cdot n_2$ *variables) i.e. both problems have the same minima.*

Proof sketch For all i, j in (7), we take all the variables $\{f_{k_1 j}, \ldots, f_{k_{w_i^p} j}\}$ that correspond to w_i^p similar pixels (with singleton weights). We then collapse each set into a single variable representing their sum $g_{ij} = \sum_{l=1}^{w_i^p} f_{k_l j}$. This can be done as their coefficients in the optimization argument $\sum_{ij} f_{ij} d_{ij}$ are the same. The w_i^p constraints of the form $\sum_j f_{k_l j} = 1$ can be replaced with a single constraint demanding $\sum_j g_{ij} = w_i^p$, without changing the space of feasible solutions. This can be done, in a similar manner, to signature Q leading to optimization problem (8). ∎

We see that when sets P and Q contain identical items it lowers the computational cost of the 1:1 matching using EMD formulation. Hence clustering similar items and replacing them with a single object is an attractive approximation to the likelihood. However this approximation degrades as the clustering becomes coarser. We can bound this error in likelihood estimation as follows:

Proposition 3.3 *Assume the ground distance* $d(p,q)$ *is a metric. Let* $P = \{(p_1,w_1^p),\ldots,(p_{n_1},w_{n_1}^p)\}, Q = \{(q_1,w_1^q),\ldots,(q_{n_2},w_{n_2}^q)\}$ *be two signatures and let* $\widehat{P}, \widehat{Q}$ *be crude versions of* P, Q *such that any object in* $\widehat{P}$ *is created by uniting objects in* P *and the same holds for* $\widehat{Q}, Q$*. Denote by* h_p, h_q *the functions mapping each object* P, Q *to its containing object in* $\widehat{P}, \widehat{Q}$*. Then:*

$$\begin{array}{l} |EMD(P,Q,d) - EMD(\widehat{P},\widehat{Q},d)| \leq \\ \sum_{i=1}^{n_1} w_i^p d(p_i, \widehat{p}_{h_p(i)}) + \sum_{i=1}^{n_2} w_i^q d(q_i, \widehat{q}_{h_q(i)}) \end{array} \quad (10)$$

In other words, the EMD approximation gap is bounded by the sum of distances between the original cluster centers and their cruder counterparts in the crude signatures. The proof of proposition 3.3 is provided in the supplementary material.

4. Noise Model

We have shown that Locally Orderless Matching attempts to explain a set P as a noisy replica of set Q, under some noise model with parameters Θ. We now turn our attention to the choice of the noise model and ways to estimate its parameters from the data. In general, any distribution can be used as a noise model. One can use prior knowledge, theoretical or empirical, about the noise to make an educated choice. In particular we consider the case of Gaussian noise for both location and appearance, assuming independence between the two, i.e. $Pr(p|q,\Theta_L,\Theta_A) = Pr(p^L|q^L,\Theta_L)\cdot Pr(p^A|q^A,\Theta_A)$.

4.1. Gaussian Noise

A Gaussian with zero mean and scalar covariance is considered for both appearance and location.

$$\begin{array}{l} Pr(p^L|q^L) \sim N(0,\Sigma_L=\sigma_L\cdot I) \\ Pr(p^A|q^A) \sim N(0,\Sigma_A=\sigma_A\cdot I) \end{array} \quad (11)$$

Denoting $\Theta=(\sigma_L,\sigma_A)$. The conditional probability is:

$$Pr(p|q,\Theta) = \frac{1}{2\pi\sigma_L^2}e^{-\frac{||p^L-q^L||_2^2}{2\sigma_L^2}}\cdot\frac{1}{(2\pi)^{k/2}\sigma_A^k}e^{-\frac{||p^A-q^A||_2^2}{2\sigma_A^2}} \quad (12)$$

Ground distance in this case is:

$$d(p,q) = \frac{1}{2\sigma_L^2}||p^L-q^L||_2^2 + \frac{1}{2\sigma_A^2}||p^A-q^A||_2^2 + C \quad (13)$$

Where $C = \frac{k+2}{2}log(2\pi) + 2log(\sigma_L) + klog(\sigma_A)$. This model is simple and intuitive, closely related to Koenderink's locally orderless image representation [1].

4.2. Parameter Estimation

Locally Orderless Matching with a Gaussian noise model of the form discussed above has two parameters σ_A and σ_L. Due to the independence assumed between appearance and location each parameter can be estimated separately using a Maximum Likelihood (ML) estimator and p,q,σ,k will be used without the superscripts A,L. Recall from propositions 3.1,3.2 that $\log Pr(P|Q,\Theta) \sim \sum_{i,j} d_{ij}f_{ij}$, where the f_{ij} providing the $1:1$ mapping, are obtained from the EMD solution. Maximum likelihood can hence be obtained by differentiating $\sum_{i,j} d_{ij}f_{ij}$ with respect to σ and comparing to zero. For $d_{ij} = d(p_i,q_j) = \frac{1}{2\sigma}||p_i-q_j||_2^2 + \frac{k}{2}log(2\pi) + klog(\sigma)$ we get:

$$\sigma^2 = \frac{1}{k}\frac{\sum_{i,j} f_{ij}||p_i-q_j||_2^2}{\sum_{i,j} f_{ij}}. \quad (14)$$

Parameter estimation can be done using a Maximization-Maximization (MM) scheme where we iterate between EMD solution and parameter update until both converge (convergence is guaranteed as both MM steps increase the likelihood). Experiments showing we can correctly estimate the noise parameters are provided in the supplementary material.

5. Locally Orderless Tracking

We are now ready to put all the pieces together. Locally Orderless Tracking applies Locally Orderless Matching to tracking. This is done in a Baysian approach using Particle-Filtering (PF) [25] where the likelihood that a certain particle has originated from the tracked object is inferred using Locally Orderless Matching. The overall algorithm is given in Algorithm 1. Specific details are provided below .

To define the conditional probability between patches $Pr(P|Q,\Theta)$ we only have to define the probabilistic noise model for single pixels $Pr(p|q,\Theta)$. The ground distance for the EMD is then defined as $d(p,q) = -log(p|q,\Theta)$ and $Pr(P|Q,\Theta)$ is obtained by solving the EMD problem.

Solving an EMD problem can be a computationally challenging task, so instead of using raw pixel values we work with superpixels. Specifically, target and candidate patches are represented by signatures which are generated from superpixels computed using TurboPixels [26] clustering built in a region-of-intrest (ROI) which supports all the particles. A signature consists of M clusters that reside in the signature support, i.e. a rectangle. Each cluster is represented by its location, i.e. geometric center of mass, and average appearance (e.g. average HSV values).

The target's state at each frame is found using PF. A signature is built for each of the N particles which are rectangular image patches and the EMD is then calculated between each of these candidate signatures $\{P_k\}_{k=1}^N$ and the target signature Q_0 with ground distances as explained above (calculated using the noise model parameters Θ). The EMD scores $\{EMD_k\}_{k=1}^N$ are then used to set particle weights according to $\pi_k = e^{-\beta\cdot EMD_k}$ and the new target state is taken to be the weighted sum over all particles. Finally, noise model parameters are updated as explained next and new particles are drawn for the next iteration of the algorithm.

Noise model parameters are updated based on the new target state found. The EMD flow between the final candidate signature and the target signature Q_0 is found providing the most probable $1:1$ matching between source and target signatures. Using this flow we estimate the noise distribution parameters Θ_{ML} according to (14). These estimated parameters are then regulated using a prior Θ_{Prior} and a moving average (MA) process before producing the

final parameters Θ_n:

$$\begin{aligned}\Theta_{MAP} &= \frac{\Theta_{ML}+\Theta_{Prior}\cdot w_{Prior}}{1+w_{Prior}} \\ \Theta_n &= (1-\alpha_{MA})\cdot\Theta_{n-1}+\alpha_{MA}\cdot\Theta_{MAP}\end{aligned} \tag{15}$$

Algorithm 1 ***Locally Orderless Tracking***

Input: Frame $I^{(n)}$, target signature $Q_0 = \{q_i, w_i^q\}_{i=1}^{M_{Q_0}}$, noise parameters $\Theta^{(n-1)}$, particle states $\{X_i^{(n)}\}_{i=1}^N$

Output: New target state $X_{Target}^{(n)}$, updated parameters $\Theta^{(n)}$, new particle states $\{X_i^{(n+1)}\}_{i=1}^N$

1. Partition ROI in $I^{(n)}$ into superpixels I_{SP}
2. For each particle $X_k^{(n)}$ do:
   (a) Build signature $P_k = \{p_i^k, w_i^{p^k}\}_{i=1}^{M_{P_k}}$ using I_{SP}
   (b) Compute ground distances using (13): $\{d_k\}_{ij} = d(p_i^k, q_j) = -log(p_i^k|q_j, \Theta^{(n-1)})$
   (c) Compute $EMD_k \leftarrow \text{EMD}(P_k, Q_0, d_k)$
   (d) Compute particle weight $\pi_k = e^{-\beta\cdot EMD_k}$
3. Normalize weights s.t $\sum_{i=1}^N \pi_i = 1$
4. Find new target position $X_{Target}^{(n)} = \sum_{i=1}^N \pi_i X_i^{(n)}$
5. Build new target signature P_T and compute EMD flow $f_{i,j} \leftarrow \text{EMD}(P_T, Q_0, d_T)$
6. Update parameters $\Theta^{(n)}$ according to (15).
7. Draw particles $\{X_i^{(n+1)}\}_{i=1}^N$ as explained in [25].

6. Experiments

This section presents experimental results. We begin with the experimental setup followed by a demonstration of the on-line adaptation capabilities. We then present qualitative and quantitative results on challenging sequences both commonly used and new comparing LOT with state-of-the-art methods.

6.1. Experimental Setup

For our experiments we use HSV color space for appearance description. Both appearance and location spaces are normalized to the range $[0, 1]$. Cluster weights are determined according to the fraction of pixels associated with them in the signature (thus ensuring a total signature weight of 1). The state vector includes position and scale i.e. $X_i = \{x_i, y_i, w_i, h_i\}$. We use $N = 250$ particles and particle weighing parameter β is set to 10 in order to better differentiate between particle scores. The noise model parameters $\Theta = \{\sigma_A, \sigma_L\}$ are initialized according to $\sigma_{A^{prior}} = 0.05, \sigma_{L^{prior}} = 0.1$. Prior weights for noise parameter updating, as explained in section 5, are initialized to $w_{\sigma_A^{prior}} = w_{\sigma_L^{prior}} = 0.25$. The ML estimator is calculated according to (14) and the MA parameter is fixed to $\alpha_{MA} = 0.3$. All parameters are kept fixed for all experiments. We use the target signature extracted in the first frame as our target template. We note that estimating the noise parameters effectively control the space of templates that can match this target template and thus can be viewed as a form of constrained model update.

In this configuration our hybrid Matlab-Mex implementation runs at ~ 1 sec per frame for a target window size of about 50x50 pixels on a standard PC.

Sequence	IVT	OAB	MIL	VTD	LOT
Dog	87	57	45.5	70	**97.4**
Shop	**36.4**	20.9	20.9	35	34.6
Girl	15.4	26.2	25	**93.4**	67.6
Human	88.8	26.2	25	64.6	**97.6**
Skating	3.8	8.8	9.8	11.5	**29.4**
Lemming	16.2	37.1	37.6	54.3	**73.8**
David	**83.1**	9.7	19.3	18.8	10
Sylv	45.7	31	73.2	**93.4**	67.6
Face	**99**	75	54.5	70.1	44.4

Table 1. Quantitative comparison, for 9 commonly used sequences, showing the percent of frames for which the PASCAL criterion was $a_0 > 0.5$. Best result are in **bold** preface. It can be seen that LOT (the proposed method) is comparable to the state-of-the-art as it gives the best results in 4 out of 9 sequences and is in second place in 2 additional sequences.

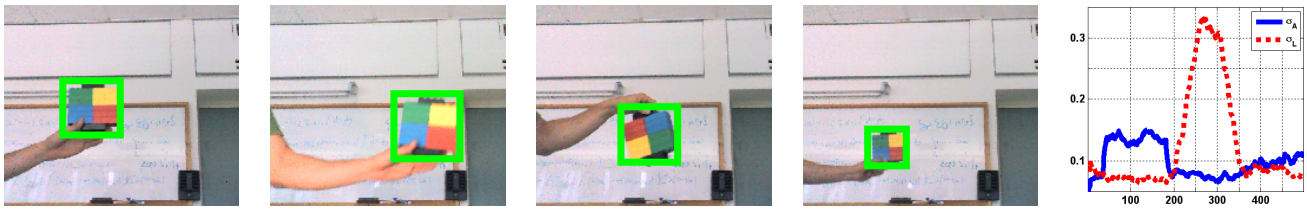
6.2. On-line Parameter Update with Toy Example

We first demonstrate the on-line noise parameter update capabilities of LOT using a 500 frame toy-example of a LEGO target subject to both appearance and localization noises. Figure 1 shows 4 sample frames from this sequence and also the behavior of the noise parameters σ_L and σ_A throughout the sequences.

The target is first subject to an illumination change. LOT detects the appearance change and increases the appearance noise parameter σ_A while maintaining perfect tracking. As the illumination returns to normal the value of σ_A decreases. Next the target is rotated about its origin. Modeling only $2D$ translation (and not rotation) this rotation is effectively localization noise albeit not a Gaussian noise. As before the target is tracked perfectly while LOT estimates and adapts the value of σ_L on-line increasing σ_L as the rotation angle increases and then decreasing σ_L as the target is rotated back. Towards the end of the sequence the algorithm correctly tracks target scale changes without altering the noise parameters which is a desired behavior.

6.3. Results for Commonly Used Sequences

We evaluate our performance on 9 challenging sequences used in recent publications [4, 5, 27, 6, 7, 12, 13]. We compare LOT's performance with 4 state-of-the-art



Frame 1 Frame 135 Frame 290 Frame 420 Parameter Estimation

Figure 1. Parameter estimation for the LEGO sequence: (Right) Noise parameter values, σ_L (Dashed-Red) and σ_A (Solid-Blue) per frame showing their on-line update.(Left) Four sample frames. First the target is illuminated with a strong light causing an appearance change handled by increasing σ_A. Next the target is rotated and since we only model $2D$ translation (w/o rotation) this creates localization noise which is handled by a large σ_L variation. Finally the target moves away from the camera causing a scale change which is correctly tracked without significant noise parameter changes.

tracking algorithms with publicly available implementations: Visual Tracking Decomposition (VTD)[7], Multiple Instance Learning (MIL)[27], Incremental Visual Tracking (IVT)[5] and Online AdaBoost (OAB)[11].

We adopt the widely used PASCAL VOC[28] criterion which quantifies both the centering accuracy as well as the scale accuracy. The criterion is $a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}$ where B_p and B_{gt} denotes the predicted and ground truth bounding boxes accordingly. Successful tracking is considered as $a_0 > 0.5$ (50%). We note that some of the sequences were re-annotated in order to provide ground truth for each frame that also accounts for scale changes disregarded in some of the original annotations.

Quantitative results are presented in Table 1 where it can be seen that LOT's performance is comparable to the state-of-the-art algorithms. LOT provides best performance in 4 out of 9 sequences (Dog, Human, Skating and Lemming) it measures up to IVT and VTD for Shop and holds second place for Girl. The remaining 3 sequences (David, Sylv and Face) are gray-scale sequences. LOT can run in both color and gray-scale (e.g. Dog), however using color appearance representation (i.e. HSV) makes gray-scale more challenging as it leaves the algorithm with only a single appearance channel. This makes coping with severe global and local illumination changes a difficult challenge and it is mainly for this reason that LOT's performance degrades on the last 3 sequences where it ranks fifth and last in David, third in Sylv and last in Face. Although VTD and IVT produce better results for some sequences, looking at the entire dataset it can be seen that no other method provides better overall performance than LOT. We believe that MIL and OAB have poorer performance mainly due to their lack of scale adaptability.

Figure 2 presents sample frames from two sequences (Dog and Skating) qualitatively showing LOT's ability to cope with difficult appearance changes such as massive scale changes and out-of-plane-rotations.

Sequence	IVT	OAB	MIL	VTD	LOT
DH	8.9	47.8	45.5	69.4	**92.3**
Shirt	0.5	66.7	32.5	79	**88.1**
Train	2.7	3.4	2.3	2.9	**69.6**
UCSDPeds	26.4	42.5	31.8	60.5	**73.9**
Boxing	7.3	18.75	18.4	21.2	**70.1**

Table 2. Quantitative comparison, for 5 new sequences, showing the percent of frames for which the PASCAL criterion was $a_0 > 0.5$. Best result are in **bold** preface. It can be seen that LOT (the proposed method) outperforms the other methods producing significantly better results on this set of challenging sequences.

6.4. Results for New Sequences

In this part we present additional results on 5 challenging new sequences. The videos include gray-scale and color examples with both static and moving cameras. The targets in these sequences are subject to many appearance changes due to deformations, pose changes, out-of-plane-rotations, massive scale changes, motion blur and illumination changes.

A quantitative comparison, based on the PASCAL criterion, between LOT and the four state-of-the-art methods is presented in Table 2. It can be seen that LOT outperforms the other tracking methods producing significantly better results for all these sequences.

Sample frames from three of the sequences are presented in Figure 3.

The first, 481 frame long, sequence shows a Down-Hill (DH) bike ride. As the rider jumps and moves in and out of shade a lot of motion blur, deformations and illumination changes are created. IVT drifts after the first jump , MIL and OAB keep tracking but eventually also drift. Only VTD and LOT are able to track the rider until the end of the sequence.

The second, 951 frame long, sequence we captured is of a T-shirt undergoing severe non-rigid deformations and motion blur. All 4 competing algorithms are unable to track the target through the severe non-rigid deformations and loose

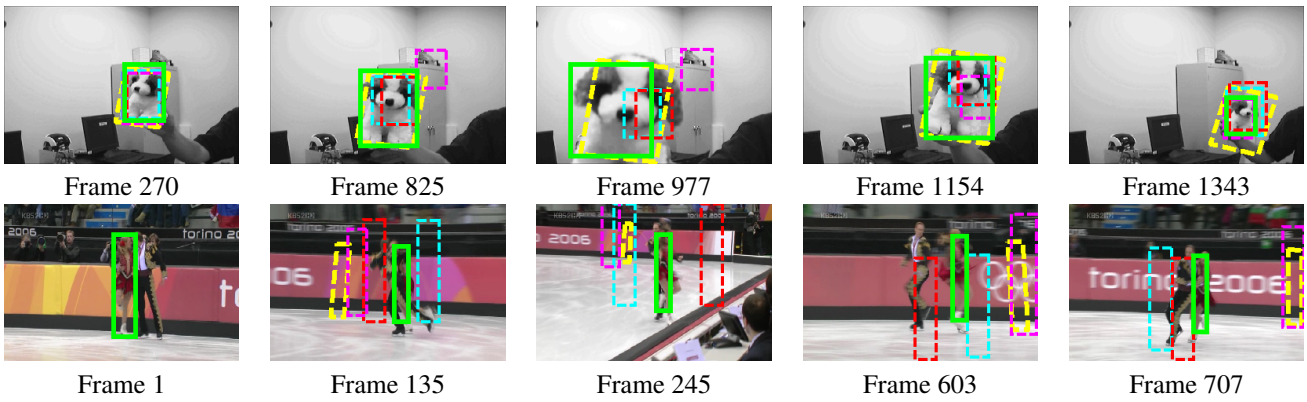


Figure 2. Sample frames from two sequences: Dog and Skating . The different algorithms are: IVT in Yellow, OAB in Cyan, MIL in Red, VTD in Magenta and LOT (The proposed algorithm) in Green.

track at some point. Only LOT with its inherent ability to explain non-rigid deformations tracks the shirt through the entire sequence.

The third, 900 frame long, sequence was taken from the PETS-2006 dataset[1]. It shows a man walking around a busy train station making many pose changes and undergoing several occlusions. Although LOT does not have an explicit mechanism for handling occlusions it can handle partial occlusions by tracking the remaining visible part of the target which often captures the full target color statistics. In this sequence the first partial occlusion occurs around frame 35 causing IVT, OAB and MIL to drift. A second occlusion at around frame 60 throws VTD off track as well. LOT is able to overcome these 2 occlusion by shrinking and matching to the remaining visible part of the target. It continues tracking the man for the entire length of the sequence while overcoming pose changes and additional occlusions.

The forth, 261 frame long, gray-scale, sequence taken from the UCSD crowd dataset[2] shows two people walking and fighting. We track both people as a single target. This crowd target undergoes non-rigid deformations as the people draw nearer and apart and as they fight with each other. Although all the methods are able track the targets location throughout most of the sequence with only minor glitches, LOT produces significantly more accurate results.

The fifth and last is a, 352 frame long, boxing sequence. At the beginning of this sequence only LOT is able to correctly track the boxer through the difficult pose changes. All methods drift between frame 200-225 due to a rapid movement followed by an occlusion however LOT is able to lock back on at frame 281 and continue tracking the target until the end of the sequence.

[1] http://www.cvg.rdg.ac.uk/PETS2006/data.html
[2] http://www.svcl.ucsd.edu/projects/peoplecnt/index.htm

7. Conclusions

Locally Orderless Tracking is a new visual tracking algorithm that estimates and adapts, on-line, to the rigidity of the tracked object. The algorithm is governed by a small set of parameters Θ that are estimated on-line allowing it to go from rigid template matching on one end to histogram-like tracking on the other, or be anywhere in between. At the heart of this framework lies Locally Orderless Matching, a new probabilistic interpretation of EMD that rigorously shows how EMD can be used to infer the likelihood that patch P is a noisy replica of patch Q with noise parameters Θ. We have shown how these noise parameters can be estimated from the data at hand and also presented results demonstrating this on-line estimation and adaptation. Finally we have shown that LOT's performance is comparable to state-of-the-art methods on a wide range of commonly used and new videos presenting superior performance in many cases.

The framework developed in this work is generic to any noise model and appearance space, future work is intended to look into different noise models and appearance representations that might be better suited for specific applications.

References

[1] J. J. Koenderink and A. J. Van Doorn, "The structure of locally orderless images," *IJCV*, 1999.

[2] B. V. Ginneken and B. M. T. H. Romeny, "Applications of locally orderless images," in *Scale-Space Theories in Computer Vision*, Springer, 1999.

[3] G.D.Hager and P.N.Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *TPAMI*, 1998.

[4] D.Ross, J.Lim, and M.H.Yang, "Adaptive probabilistic visual tracking with incremental subspace update," *ECCV*, 2004.

Figure 3. Sample frames from three sequences: DH, Shirt, and Boxing. The different algorithms are: IVT in Yellow, OAB in Cyan, MIL in Red, VTD in Magenta and LOT (The proposed algorithm) in Green.

[5] D.Ross, J.Lim, R.S.Lin, and M.H.Yang, "Incremental learning for robust visual tracking," *IJCV*, 2007.

[6] X.Mei, H.Ling, Y.Wu, E.Blasch, and L.Bai, "Minimum error bounded efficient l1 tracker with occlusion detection," *ICCV*, 2011.

[7] J.Kwon and K.M.Lee, "Visual tracking decomposition," *CVPR*, 2010.

[8] D.Comaniciu, "Bayesian kernel tracking," *DAGM*, 2002.

[9] M. P. H.Bischof, "Hough-based tracking of non-rigid objects," *ICCV*, 2011.

[10] S. Avidan, "Ensemble tracking," *CVPR*, 2005.

[11] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via online boosting," *BMVC*, 2006.

[12] S.Wang, H.Lu, F.Yang, and M.H.Yang, "Superpixel tracking," *ICCV*, 2011.

[13] J.Santner, C.Leistner, A.Saffari, T.Pock, and H.Bischof, "Prost:parallel robust online simple tracking," *CVPR*, 2010.

[14] A.Elgammal, R.Duraiswami, and L.S.Davis, "Probabilistic tracking in joint feature-spatial spaces," *CVPR*, 2003.

[15] S. Peleg, M. Werman, and H. Rom, "A unified approach to the change of resolution: Space and gray-level," *TPAMI*, 1989.

[16] C. Y.Rubner and L.J.Guibas, "The earth mover's distance as a metric for image retrieval," *IJCV*, 2000.

[17] E.Levina and P.Bickel, "The earth mover's distance is the mallows distance: some insights from statistics," *ICCV*, 2001.

[18] Q. Zhao, Z. Yang, and H. Tao, "Differential earth mover's distance with its applications to visual tracking," *PAMI*, p. 7, 2010.

[19] X. Ren and J. Malik, "Learning a classification model for segmentation," *ICCV*, 2003.

[20] D. Hoiem, A. Efros, and M. Hebert, "Geometric context from a single image," *ICCV*, 2005.

[21] X. He, R. Zemel, and D. Ray, "Learning and incorporating top-down cues in image segmentation," *ECCV*, 2006.

[22] S. S. Boltz, F. Nielsen, "Earth mover distance on superpixels," *ICIP*, 2010.

[23] A. Yilmaz, O. Javed, and M. Shah., "Object tracking: A survey," *ACM. Comp. Survey*, vol. 38(4), 2006.

[24] I.Heller and C.B.Gh.Tompkins, "An extension of a theorem of dantzig's," *Linear Inequalities and Related Systems, Annals of Mathematics Studies, 38, Princeton (NJ)*, pp. 247–254, 1956.

[25] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *IJCV*, 1998.

[26] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows," *TPAMI*, 2009.

[27] B.Babenko, M.H.Yang, and S.Belongie, "Visual tracking with online multiple instance learning," *CVPR*, 2009.

[28] M.Everingham, L.J.V.Gool, C. Williams, J.M.Winn, and A.Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, 2010.