Granularity and Elasticity Adaptation in Visual Tracking

Ming Yang, Ying Wu
Dept. of EECS, Northwestern Univ.
2145 Sheridan Rd., Evanston, IL 60201, USA
{mya671,yingwu}@ece.northwestern.edu

Abstract

The observation models in tracking algorithms are critical to both tracking performance and applicable scenarios but are often simplified to focus on fixed level of certain target properties such as appearances and structures. In this paper, we propose a unified tracking paradigm in which targets are represented by Markov random fields of interest regions and introduce a new way to adapt observation models by automatically tuning the feature granularity and model elasticity, i.e. *the abstraction level of features and the model's degree of flexibility to tolerate deformations. Specifically, we employ a multi-scale scheme to extract features from interest regions and adjust the parameters of the potential functions of the MRF model to maximize the likelihoods of tracking results. Experiments demonstrate the method can estimate translation, scaling and rotation and deal with deformation, partial occlusions, and camouflage objects within this unified framework.*

1. Introduction

Visual object tracking is the core task in motion analysis and crucial to many applications. The most critical factor determining tracking performance is the matching criterion, also known as observation model or likelihood model, which defines what trackers are following. However, matching is largely simplified in most tracking methods and may only focus on certain characteristics of targets, for example, the existences of certain local visual patterns or coherence with certain overall feature statistics in appearance-based tracking. Consequently, successful tracking methods for certain type of targets may not adapt to other targets easily. Therefore, for generally applicable trackers, matching need to be flexible for distinctive targets and adaptive with respect to target variations.

Designing generally applicable trackers is extremely challenging, if not impossible, mainly due to enormous variabilities of targets and their unpredictable changes in practice. Actually, it is even hard to handle targets with rotation and scale changes in a unified way if they also experience different degrees of deformations and partial occlusion, since the observation models of targets may not be able to adapt to all these factors simultaneously. In order to advance towards designing more general trackers, adaptation of more aspects of observation models need to be introduced and incorporated in a unified framework.

Specifically, for appearance-based tracking, there are two key aspects in designing observation models: what is the abstraction level of features, and how to take into account the geometrical structures of targets. For example, in two extreme cases, the template matching method [13] uses local pixel intensities as features and employs sum of squared differences (SSD) as the matching criterion that enforces rigid geometrical relations among pixels, so it is suitable for small and rigid targets but vulnerable to partial occlusions and deformations. On the other hand, kernel-based tracking algorithms [8, 7, 9] represent targets by weighted histograms that delineate the overall statistics of targets' appearances and largely ignore their geometrical layouts. Therefore these algorithms can deal with non-rigid targets with sufficient sizes but are insensitive to some motion parameters. In between of these two extreme cases, many other algorithms, such as "super pixels" [26, 24] or "bag-of-patches" approaches [3, 1, 25, 23], extract features from some regions of interest on targets and consider their geometrical relations to different degrees.

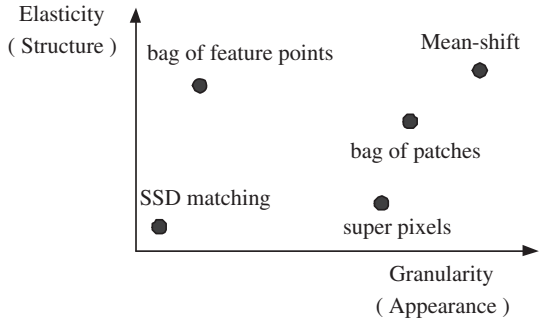

Figure 1. Illustration of different tracking approaches in terms of their relative granularity and elasticity.

We refer the two aforementioned dimensions as the fea-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

ture *granularity* and model *elasticity*. Granularity is a measure of descriptions of components that make up an object. We use the feature granularity to indicate the abstraction level of features, *e.g.* whether features describe attributes of a pixel, a blob region or a whole object. Elasticity refers to the degree of flexibility. Here we use the model elasticity to indicate the ability that the model tolerates geometrical changes among components, *e.g.* whether a model allows deformations inside targets or not. The feature granularity focuses on the target appearance and the model elasticity puts emphasis on its structure. Some typical tracking approaches are illustrated qualitatively in Fig. 1 in terms of their relative feature granularity and model elasticity.

Human also perceive different objects at different granularity levels [12]. For objects full of textures but without clear structures, human eyes may focus on their local appearance characteristics. For objects composed by several parts, both the appearances of the parts and their structures may attract the attention. In addition, as the scales of objects change or deformation/partial occlusion occurs, the perception of target structure and local appearance may also change. Inspired by these observations, a natural question is whether trackers can adapt the feature granularity and model elasticity in their observation models.

In this paper, we propose a unified tracking paradigm in which targets are represented by Markov random fields (MRF) of a set of interest regions where the feature granularity and model elasticity can be explicitly adapted with respect to targets' appearances during tracking. The feature vectors that delineate the local appearances of interest regions are extracted in a multi-scale manner. Thus, the scale ratio between the patch sizes that are used to extract feature vectors and the characteristic scales of interest regions specifies the feature granularity. On the other side, the geometrical relations among the interest regions, *i.e.* the structures of targets, are modelled in the pair-site potential functions whose parameters control the elasticity of the model. Thus, by updating the scale ratio and the parameters in the potential functions to maximize the joint likelihood of the MRF, the tracker adaptively balances the requirements of consistency with the local appearances and structures of targets.

The main contributions of the paper are on two-fold. First, the proposed tracking paradigm can be viewed as a unification of many previous tracking algorithms in the sense of how to organize appearance-based features in target observation models. Second, the adaptation of feature granularity and model elasticity in this paradigm exhibits a new way to update observation models to handle dynamical targets. The proposed method can estimate multiple motion parameters including translation, rotation and scaling, and handle partial occlusion, deformable targets and camouflage objects within the unified framework as demonstrated by extensive experiments.

2. Related work

Visual tracking has received intensive research efforts for decades and different tracking algorithms may have quite different applicable scenarios due to the use of different target observation models. Generally, feature vectors are extracted from hypothesis regions and are evaluated against target observation models to find the optimal match. For appearance-based tracking, the feature vectors may concisely abstract the properties of targets at different granularity levels, *e.g.* pixel intensity patterns [18, 13, 4, 2] or filter bank responses [16, 11], a set of feature points or interest regions [3, 1, 26, 24, 25, 23], or statistics of entire objects [8, 7, 9]. On the other hand, the geometrical structures of targets can be enforced fairly strictly in an intensity template [13], a linear subspace [4, 15], a classifier [2], or "super pixels" templates [26, 24], or loosely modelled in kernel functions [8, 7, 9] or a "bag of patches" where targets are located with the confidence or occupance maps [3, 11, 1, 25, 23] based on matching interest regions.

Most of these tracking algorithms involve observation models with pre-defined degree of focus on targets' local appearances or structures. Although the observation models can be updated by latest tracking results [16, 15], online classification [3, 11], or selection of different cues [7] and feature points [22], the feature granularity and model elasticity remain roughly at the same level. In contrast, in our approach, we represent a target by an MRF of interest regions where the feature granularity and model elasticity are able to explicitly adapt to distinctive targets by extracting features with different scale ratios and tuning the potential functions. Note, the proposed method is different from some recent work [25, 23, 22] where target is represented by a constellation of fixed-size (11×11) intensity patches extracted at Harris corners [25], or a bag of maximally stable extremal regions (MSER) [20], or an attributed relational graph of SIFT features [22].

The constellation model of feature points has been deeply investigated and very successful in object recognition and categorization [17, 10, 5, 6]. Comprehensive survey of all graph-based object representations in computer vision literature is out of the scope of the paper.

3. Target observation model

In this paper, we propose a unified tracking paradigm where the target is represented by an MRF model of interest regions, and the feature granularity and the model elasticity can be explicitly modelled in a parametric way. In this section, we first introduce the general tracking paradigm and then describe the specific interest regions and MRF formulation in our implementation.

3.1. A unified tracking paradigm

Given the target initialization, we construct an MRF based on the interest regions within the target. The hidden

variables $\mathbf{X} = \{\mathbf{x}_i\}$ in the MRF are the parameters of the interest regions on the target, and the observable variables are the parameters $\mathbf{Y} = \{\mathbf{y}_i\}$ of detected interest regions in every frame. The adjacent interest regions are linked in pair-wise cliques that encode their relative geometrical relations, as shown in Fig. 2. Then, by matching features extracted from the interest regions in successive frames, the motion of the targets can be first coarsely estimated based on the motion of the interest regions. Afterwards, we refine the target's motion parameters by searching the maximum a posteriori (MAP) estimate $P(\mathbf{X}^*|\mathbf{Y})$. We employ the scale ratio between the sizes of image patches to extract features and the characteristic scales of interest regions to model the feature granularity. The elasticity of the model is controlled by the parameters in the potential functions. Assuming the tracking results are true realizations of the MRF, we adapt the granularity and elasticity to maximize the joint probability $P(\mathbf{X}^*)$. The entire paradigm is summarized in Fig. 3.

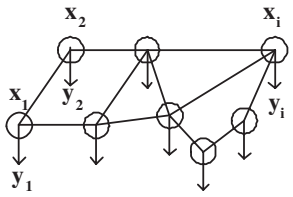

Figure 2. Illustration of the MRF model.

With different types of interest regions and strategies in extracting features, the MRF-based observation model in this tracking paradigm can substantialize to different observation models. For example, if we regard each pixel as an interest region and enforce strict geometrical relations among the pixels, this model degenerates to template tracking, or if the entire object is an interest region and features are kernel-weighted histograms, then it turns to kernel-based tracking. Additionally, "bag-of-patches" method can be well categorized into the paradigm if no geometrical constraints are enforced in the MRF and the motions of targets are estimated from the confidence map or probabilistic occupance map generated from interest region matching or outputs of classifiers.

3.2. Interest region detection

For interest regions in the MRF, salient image patches that are stable in affine transforms are preferable since their motion parameters can be explicitly estimated. There are many successful affine region detection methods [20], and we select Harris-Laplace interest regions in our implementation mainly due to its computational efficiency and the ability to yield rich candidate regions.

The Harris-Laplace interest point detector [14, 19] extracts points that are both local maxima of the Harris cornerness measure in spatial domain and maxima of the normalized Laplacian in scale space. The cornerness is measured based on the second moment matrix μ of the image gradient distribution in a neighborhood of a pixel $\{u, v\}$, as

$$\mu(\{u,v\}, s_I, s_D) = s_D^2 g(s_I) \otimes \begin{pmatrix} L_u^2(\{u,v\}, s_D) & L_u L_v(\{u,v\}, s_D) \\ L_u L_v(\{u,v\}, s_D) & L_v^2(\{u,v\}, s_D) \end{pmatrix}, \quad (1)$$

where $L_u(\{u,v\}, s_D)$ and $L_v(\{u,v\}, s_D)$ are image gradients after smoothed by a Gaussian kernel with variance s_D, *a.k.a.* the derivation scale [19], and $g(s_I)$ indicates the Gaussian kernel to integrate the gradients whose variance s_I is referred as the integration scale or the characteristic scale [19] of this point. The two eigenvalues $\lambda_1 \geq \lambda_2$ of μ characterize the pixel intensity distributions in the neighborhood. Two large eigenvalues imply the motion of the image patch surrounding this pixel may be phenomenal in all directions [21], thus it is a stable corner. Each Harris corner can be delineated by an ellipse region R centered at $\{u, v\}$ with the characteristic scale s_I and a shape matrix $\bar{\mu}$ that are normalized by the larger eigenvalue λ_1.

After extracting the ellipses $R = \{u, v, s_I, \bar{\mu}\}$ whose centers are Harris corners, we calculate the normalized Laplacian for those nested ellipses, that is, $R^{'}$ and R are nested if $R^{'} \subset R$. Note, the centers are not necessarily the same for the nested ellipses. The regions that are local maxima of the normalized Laplacian $s_D^2|L_{uu}(\{u,v\}, s_D) + L_{vv}(\{u,v\}, s_D)|$ are selected as the detected interest regions $\{R_1^t, \cdots, R_{M^t}^t\}$ where M^t denotes the number of regions detected at frame t.

Please refer to [19] for details about Harris-Laplace interest point detector. In [19], the location and shape of an interest region are iteratively refined in order to reflect the gradient distributions more accurately. As there is no guarantee of the convergence and the computation load is not affordable for tracking, we do not refine the interest regions.

3.3. MRF model formulation

Given the detected interest regions $\{R_1^0, \cdots, R_{M^0}^0\}$ within the initial target at frame $t = 0$, we build the MRF including the hidden sites $\mathbf{x}_i = \{u_i, v_i, s_{Ii}, \bar{\mu}_i\}$ that correspond to R_i^0 and incorporate the target's motion parameters in the pair-wise potential functions.

The initial interest regions $\{R_1^0, \cdots, R_{M^0}^0\}$ are regarded as a true realization of the MRF and denoted as $\{\mathbf{x}_1^0, \cdots, \mathbf{x}_{M^0}^0\}$. Then, the joint probability $P(\mathbf{X}) = P(\mathbf{x}_1, \cdots \mathbf{x}_{M_0})$ is expressed by the Gibbs energy defined over pair-wise clique set C, as

$$P(\mathbf{X}) = \frac{1}{Z} \exp - \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in C} V(\mathbf{x}_i, \mathbf{x}_j), \quad (2)$$

where Z is the partition function and V is the pair-wise potential function. $(\mathbf{x}_i, \mathbf{x}_j)$ is a pair-wise clique if the corresponding interest regions overlap. The higher order cliques

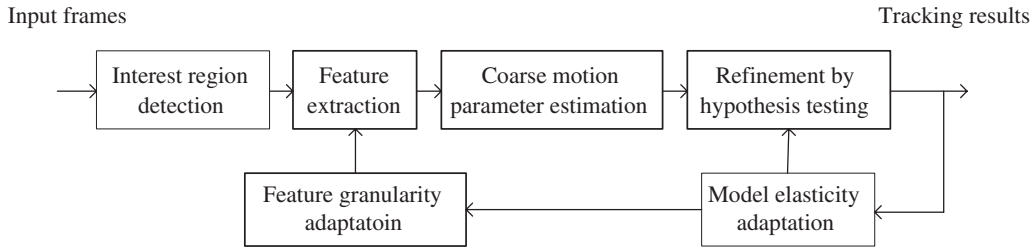

Figure 3. The proposed unified tracking paradigm.

and the dependencies among cliques with common interest regions are ignored to enable the problem tractable.

It is open and flexible to define the potential function V to model the relative geometrical relation between two interest regions. To allow rotation and scaling of targets, in V we only involve the difference of the angle θ_{ij}^t between $\mathbf{x}_i^t$ and $\mathbf{x}_j^t$ at frame t against the reference angle θ_{ij}^0 between $\mathbf{x}_i^0$ and $\mathbf{x}_j^0$, and the target's current rotation angle $\Delta\theta^t$, as

$$V(\mathbf{x}_i^t, \mathbf{x}_j^t) = \frac{(\theta_{ij}^t - \theta_{ij}^0 - \Delta\theta^t)^2}{2\sigma^2}, \tag{3}$$

where σ is the assumptive variance of angle differences $\Delta\theta_{ij}^t = \theta_{ij}^t - \theta_{ij}^0$, which can control the elasticity of the MRF, *i.e.* how rigid the relative geometrical relations among interest regions are enforced. The angle θ_{ij}^t between two adjacent interest regions is calculated with the link connecting their centers, *i.e.* $\theta_{ij}^t = \arctan(\frac{v_i^t - v_j^t}{u_i^t - u_j^t})$. With these definitions, the partition function Z can be explicitly expressed as $Z = (\sqrt{2\pi}\sigma)^{|C|}$ where $|C|$ is the number of pair-wise cliques. An example of MRF model is illustrated in Fig. 4 where the interest regions are drawn as yellow ellipses and the centers of those that are neighbors are linked with red lines.

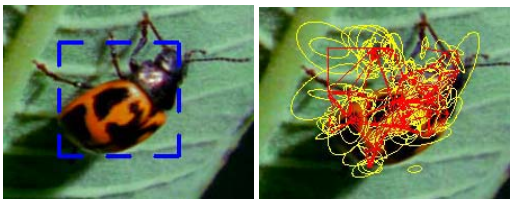

Figure 4. An example of the MRF model initialization.

Since histograms are generic and rotation invariant, for each interest region, we extract a histogram of certain cue to describe its appearance. For a Harris-Laplace interest point, although the characteristic scale s_I is available, but how large area around the point should be used to extract the features to insure good matching can not be determined before tracking. Thus, we utilize a scalar r to specify the scale ratio between the size of image patch used to extract the histogram and the characteristic scale s_I. For each $\mathbf{x}_i$, $H(r\mathbf{x}_i)$ represents the histogram extracted from the ellipse with the length of the major axis equal to rs_I. Therefore, the ratio r controls the feature granularity.

For an observation $\mathbf{y}_i^t$ of $\mathbf{x}_i$, we define the likelihood of individual interest region based on the Bahattachaya coefficient ρ between the corresponding histograms, as

$$P(\mathbf{y}_i^t|\mathbf{x}_i) = \exp(1 - \rho(H(r\mathbf{x}_i^0), H(r\mathbf{y}_i^t))). \tag{4}$$

Fixed r may not be appropriate for all tracking scenarios, so r need to be adjusted during tracking.

4. Motion estimation

We estimate the motion parameters of the target with two steps. First, the interest regions detected in current frame are matched with the initial regions in the MRF so as to coarsely estimate target's motion parameters, *i.e.* translation, scale and rotation angle, which mainly relies on the resemblance of appearance. Then, a few more motion parameters are sampled guided by the coarse estimates. The hypothesis that yields the highest joint posterior probability of the MRF is regarded as the tracking result, which takes both appearance and structures into consideration.

4.1. Coarse motion estimation

For every incoming frame, we perform Harris-Laplace interest points detector to locate the interest regions $\{R_1^t, \cdots, R_{M^t}^t\}$ at current frame in an enlarged region surrounding the previous tracking result. If one interest region is matched to an initial interest region $\mathbf{x}_i^0$, we regard it as an observation of the hidden site $\mathbf{x}_i$ and denote it by $\mathbf{y}_i^t$. The matching can be achieved by a classifier [3, 11], instead, we directly threshold the Bhattacharya coefficient ρ with the scale ratio r by a threshold T, as

$$\rho(H(r\mathbf{x}_i^0), H(r\mathbf{y}_i^t)) > T. \tag{5}$$

This matching is not necessarily a one-on-one mapping.

Incremental estimation of the motion parameters of targets, especially for the rotation angle, is not reliable since the estimation error could be accumulated. Thus, we estimate the target motion $\Delta u^t, \Delta v^t, \Delta s^t, \Delta\theta^t$ with respect to the target initialization. These motion parameters are first coarsely estimated by $\Delta u_i^t, \Delta v_i^t, \Delta s_{ij}^t, \Delta\theta_{ij}^t$ of individual observations $\mathbf{y}_i^t$ and each pair of $\mathbf{y}_i^t$ and $\mathbf{y}_j^t$ within a clique.

The translations $\Delta u_i^t = (u_i^t - u_i^0)$ and $\Delta v_i^t = (v_i^t - v_i^0)$ are cast in a 2D histogram. The scale factor and the rotation angle are estimated through 1D histogram of those of the detected pair-wise cliques $(\mathbf{y}_i^t, \mathbf{y}_j^t)$,

$$\Delta s_{ij}^t = \frac{|\{u_i^t, v_i^t\} - \{u_j^t, v_j^t\}|}{|\{u_i^0, v_i^0\} - \{u_j^0, v_j^0\}|}, \tag{6}$$

$$\Delta\theta_{ij}^t = \theta_{ij}^t - \theta_{ij}^0. \tag{7}$$

The modes of the distributions of these motion parameters present coarse motion estimation for the target, *i.e.* $\Delta\bar{u}^t, \Delta\bar{v}^t, \Delta\bar{s}^t, \Delta\bar{\theta}^t$. The histograms of interest regions' motion parameters are similar as the confidence map or occupance map used in "bag-of-patches" approaches and the geometrical relations among the interest regions have not been taken into account. Then, we employ these rough estimates to guide fine sampling of target motions and evaluate the posteriors to refine the motion estimation.

4.2. Motion parameter refinement

As the interest region detection and matching may contain errors, we further refine the coarse motion estimates $\Delta\bar{u}^t, \Delta\bar{v}^t, \Delta\bar{s}^t, \Delta\bar{\theta}^t$ by sampling a few more motion parameters around them and evaluating these hypotheses.

Given the observations $\mathbf{Y} = \{\mathbf{y_1}, \cdots, \mathbf{y_M}\}$ within a hypothesis target region, the MAP estimation $\mathbf{X}^* = \operatorname{argmax} P(\mathbf{X}|\mathbf{Y})$ presents the upper bound of the posterior of these observations $\mathbf{Y}$. With the Markovian properties and the field model structure $P(\mathbf{y}_i|\mathbf{X}) = P(\mathbf{y}_i|\mathbf{x}_i)$, the joint posterior can be expressed as

$$\begin{aligned} P(\mathbf{X}|\mathbf{Y}) &\propto P(\mathbf{Y}|\mathbf{X})P(\mathbf{X}) \\ &= P(\mathbf{X})\prod_i P(\mathbf{y}_i|\mathbf{x}_i). \end{aligned} \tag{8}$$

The joint probability $P(\mathbf{X})$ is calculated with Eq. 2 and Eq. 3 that utilize the hypothesis $\Delta\bar{\theta}^t$ as a parameter. The likelihood of individual interest region $P(\mathbf{y}_i|\mathbf{x}_i)$ is defined in Eq. 4. Then, the hypothesis whose optimal labelling $\mathbf{X}^*$ yields the highest posterior $P(\mathbf{X}^*|\mathbf{Y})$ is regarded as the tracking result.

5. Granularity and elasticity adaptation

In calculating $P(\mathbf{y}_i|\mathbf{x}_i)$ with Eq. 4 and $P(\mathbf{X})$ with Eq. 2 and 3, the scale ratio r to control the feature granularity and σ to control the elasticity of the MRF play important role in interest region matching and MAP estimation. Pre-defined fixed r and σ are not likely to assure good matching for different targets and challenging situations such as partial occlusions and camouflage objects nearby. Thus, we adapt them in every frame to maximize the posteriors of tracking results. The updated parameters r^t and σ^t at frame t are used in motion estimation at next frame $t+1$.

5.1. Feature granularity adaptation

We update the scale ratio by searching $r^t + \Delta r$ until local maximum of $P(\mathbf{Y^t}|\mathbf{X^t}^*) = \prod_i P(\mathbf{y}_i^t|\mathbf{x}^{t*}_i)$. Note, here locations and shapes of $\mathbf{y}_i^t$ and $\mathbf{x}^{t*}_i$ are given, only r^t affects $P(\mathbf{Y^t}|\mathbf{X^t}^*)$. This is equivalent to maximize the sum of the Bhattacharya coefficients of all observed interest regions $\mathbf{y}_i^t$ in the tracked target, as

$$r^{t*} = \operatorname{argmax} \sum_i \rho(H(r^t\mathbf{x}_i^0), H(r^t\mathbf{y}_i^t))). \tag{9}$$

The histograms $H(r^t\mathbf{x}_i^0)$ are pre-calculated and stored at tracking initialization. To reduce the computation overhead of adaptation, we perform local gradient search around $r^t \pm \Delta r$ with $r^0 = 2$ and $\Delta r = 0.1$ in our experiments. Thus, the feature granularity is updated according to the appearance changes. If the target is rigid and stable, good matching can be obtained with large ratio r. If partial occlusion or deformation happen, small r may be appropriate.

5.2. Model elasticity adaptation

The parameter σ in the pair-site potential functions controls the elasticity of the MRF. To enable σ match the degree of deformation of the target, we solve it by maximize the likelihood of the current tracking result, as

$$\frac{\partial \ln P(\mathbf{X}^{t*}|\sigma)}{\partial\sigma} = 0. \tag{10}$$

Plug in the partition function Z and the potential energy in Eq. 3 to $P(\mathbf{X}_t^*|\sigma)$, we have

$$\begin{aligned} P(\mathbf{X}_t^*|\sigma) &= \frac{1}{(\sqrt{2\pi}\sigma)^{|C|}} \exp - \sum_{(\mathbf{x}_i^{t*}, \mathbf{x}_j^{t*}) \in C} V(\mathbf{x}_i^{t*}, \mathbf{x}_j^{t*}) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^{|C|}} \exp - \sum_{(\mathbf{x}_i^{t*}, \mathbf{x}_j^{t*}) \in C} \frac{(\Delta\theta^{t*}_{ij} - \Delta\theta^{t*})^2}{2\sigma^2}. \end{aligned}$$

Solving Eq. 10, we obtain the assumptive variance σ^t of angle differences given the current tracking result,

$$\sigma^t = \frac{1}{|C|} \sum_{(\mathbf{x}^{t*}_i, \mathbf{x}^{t*}_j) \in C} (\Delta\theta^{t*}_{ij} - \Delta\theta^{t*})^2. \tag{11}$$

The optimal σ^t is the variance of the observed angle differences. So, if the relative geometrical relations of the detected interest regions are stable, σ^t is small, on the other hand, σ^t increases when deformations occur.

6. Experiments

We evaluate the proposed tracking algorithm for a variety of real-world sequences that present deformations, partial occlusions, and camouflage objects. In the Harris-Laplace interest point detector, up to 12 different integration scales are tested depending on the size of the target.

The features used to match the interest regions are 2D histograms in normalized-RG space with 24×24 bins and the corresponding matching threshold for the Bhattacharya coefficient is set to $T = 0.75$. The proposed tracker is implemented with C++ which runs at 2-10 frame per second on a Pentium-IV 3GHz desktop. The computation load is jointly determined by the number of scales in the interest region detector and the number of interest regions detected.

To exhibit the generality of the proposed method, for different sequences, we compare the performance with three trackers: a Mean-shift tracker that also employs 2D histograms in normalized-RG space with 24×24 bins, a template tracker where the image regions are normalized to grey-level patches and compared with SSD, and a "bag-of-patches" tracker using the same set of interest regions but ignoring their geometrical relations. Although these 3 trackers can deal with different kinds of tracking scenarios, we demonstrate the proposed method can overcome some difficulties to them within the unified tracking paradigm.

6.1. Illustration of tracking results

The tracking results are displayed in three rows in Fig.5. At the first row, the initialization of the MRF model is shown in the first image where the cliques are drawn with red lines, and followed by the interest region detection results where the matched regions are drawn as yellow ellipses while the non-matched ones are light blue ellipses. Note the length of the major axis in drawing is the product of the scale ratio r^t and the interest region's characteristic scale s_I. Our tracking results are illustrated at the second row where the target is indicated by a blue dash bounding box and the pixels covered by matched interest regions are highlighted with red boundaries. The comparison tracking results are shown at the third row.

In sequence [`Sidewalk`], the size of target is small which is suitable for the template tracker. However, when a bicycler is passing by the pedestrian from frame 140, the template tracker is easily distracted, shown in the third row of Fig.5. In our tracker, as the interest regions on the upper body of the pedestrian remain stable, the tracker can get along with the distractions.

6.2. Partial occlusions

The sequence [`Face`] first used in [1] presents different degrees of partial occlusions. Large scale ratio r may jeopardize the interest region matching when partial occlusions occur. From Fig. 6 we can observe that in our method the scale ratio r^t is adapted to follow the changing of degree of occlusions. r^t decreases to about 1.2 at frame 285 and increases to 3 when the book moves away. For the mean-shift tracker, when partial occlusion happens, the scale estimation is no longer reliable and can hardly recover. Some representative frames are shown in Fig. 7.

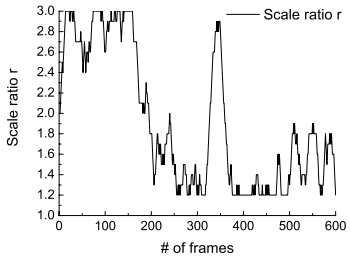

Figure 6. Scale ratio r^t for sequence [`Face`].

6.3. Deformable object

For sequence [`Cock fight`], when the target cock experiences large deformation around frame 240, σ^t in the potential functions increases considerably, as shown in Fig. 8. This means the structure or the relative geometrical relations among the interest regions are largely ignored, thus, the target is located mainly by matching its appearance. When the cock pauses fighting at frame 250, its structure helps the proposed tracker to locate the target and estimate the scale more accurately than the Mean-shift tracker.

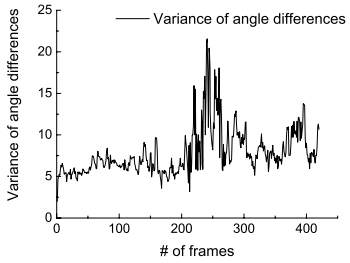

Figure 8. σ^t for sequence [`Cock fight`].

6.4. Camouflage object

If the appearance of the target is distinctive in the scene, "bag-of-patches" approaches may work well, however, they are usually vulnerable when camouflages, *i.e.* similar or even identical objects, present close to the target. As shown in Fig. 10, when the camouflage package moves close to the target from frame 640, the scale estimation in the pure "bag-of-patches" tracker becomes unstable and it gradually drifts to the wrong target. In our approach, though interest regions detected on the camouflage package have similar appearances, they are excluded since their relative positions are not consistent with the MRF model.

7. Conclusions

In this paper, we introduce a new perspective of adapting target observation models in terms of the feature granularity and model elasticity in a unified tracking paradigm, where targets are represented by MRFs of interest regions. By employing a multi-scale scheme to extract features from interest regions and adjusting the parameters that regulate the

Figure 5. Tracking [Sidewalk] for frame #1, 140, 145, 152 and 163, (1st row) initialization and interest region detection, (2nd row) the proposed tracker (3nd) the template tracker.

Figure 7. Tracking [Face] for frame #1, 285, 345, 585 and 599, (1st row) initialization and interest region detection, (2nd row) the proposed tracker (3nd) the Mean-shift tracker.

target geometrical layout, the proposed method automatically tunes the observation model's focus on target's appearances and structures. Furthermore, the proposed tracking paradigm is flexible to incorporate different interest regions and features. Future work will include investigation about how to adapt the feature granularity of individual interest regions and the potential functions for each clique.

Acknowledgments

This work was supported in part by NSF Grants IIS-0347877 and IIS-0308222.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *CVPR'06*, volume 1, pages 798 – 805, June 17-22, 2006. 1, 2, 6

[2] S. Avidan. Subset selection for efficient SVM tracking. In *CVPR'03*, volume 1, pages 85 – 92, June 16-22, 2003. 2

[3] S. Avidan. Ensemble tracking. In *CVPR'05*, volume 2, pages 494 – 501, June 20-25, 2005. 1, 2, 4

[4] M. J. Black and A. D.Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *ECCV'96*, pages 329–342, Apr. 1996. 2

[5] G. Carneiro and A. D.Jepson. The distinctiveness, detectability, and robustness of local image features. In *CVPR'05*, volume 2, pages 296 – 301, June 20-25, 2005. 2

[6] G. Carneiro and A. Jepson. Flexible spatial configuration of local image features. *IEEE Trans. Pattern Anal. Machine Intell.*, 29(12):2089 – 2104, Dec. 2007. 2

[7] R. T. Collins and Y. Liu. On-line selection of discriminative tracking features. In *ICCV'03*, volume 1, pages 346–352, 2003. 1, 2

[8] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR'00*, volume 2, pages 142–149, June 13-15, 2000. 1, 2

[9] G. D.Hager, M. Dewan, and C. V. Stewart. Multiple kernel tracking with ssd. In *CVPR'04*, volume 1, pages 790 – 797, 2004. 1, 2

Figure 9. Tracking [Cock fight] for frame #1, 229, 241, 250 and 410, (1st row) initialization and interest region detection, (2nd row) the proposed tracker (3nd) the Mean-shift tracker.

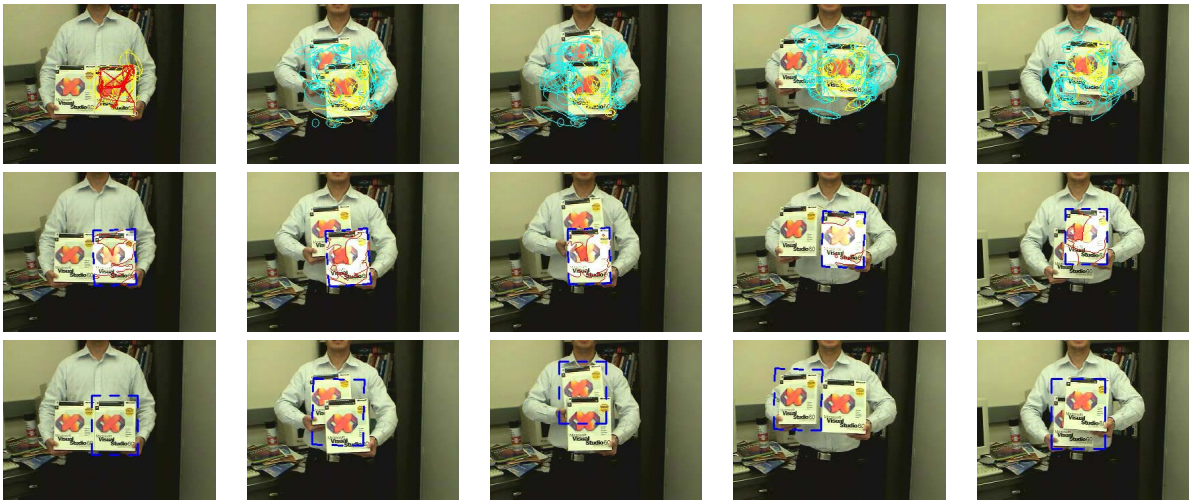

Figure 10. Tracking [Package] for frame #1, 640, 702, 740 and 792, (1st row) initialization and interest region detection, (2nd row) the proposed tracker (3nd) the "bag-of-patches" tracker.

[10] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR'03*, volume 2, pages 264 – 271, 2003. 2

[11] H. Grabner and H. Bischof. On-line boosting and vision. In *CVPR'06*, volume 1, pages 260 – 267, June 17-22, 2006. 2, 4

[12] C. Guo, S.-C. Zhu, and Y. N. Wu. Towards a mathematical theory of primal sketch and sketchability. In *ICCV'03*, volume 2, pages 1228 – 1235, Oct. 13-16, 2003. 2

[13] G. Hager and P. Belhumeur. Real-time tracking of image regions with changes in geoemtry and illumination. In *CVPR'96*, pages 403–410, June 18-20, 1996. 1, 2

[14] C. Harris and M. Stephens. A combined corner and edge detector. In *ALVEY Vision Conference*, pages 147 – 151, 1998. 3

[15] J. Ho, K.-C. Lee, M.-H. Yang, and D. Kriegman. Visual tracking using learned linear subspace. In *CVPR'04*, volume 1, pages 782–789, Jun.27-Jul.2 2004. 2

[16] A. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. In *CVPR'01*, volume 1, pages 415–422, Dec. 8-14, 2001. 2

[17] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV'99*, volume 2, pages 1150 – 1157, 20-27, 1999. 2

[18] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *DARPA Image Understanding Workshop*, pages 121 – 130, Apr. 1981. 2

[19] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV'02*, pages 128–142, May.27-Jun.2, 2002. 3

[20] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *Int'l Journal of Computer Vision*, 65(1-2):43 – 72, Nov. 2005. 2, 3

[21] J. Shi and C. Tomasi. Good features to track. In *CVPR'94*, pages 593 – 600, June20 - 24, 1994. 3

[22] F. Tang and H. Tao. Object tracking with dynamic feature graph. In *VS-PETS'05*, pages 25 – 32, Oct. 15-16, 2005. 2

[23] S. Tran and L. Davis. Robust object tracking with regional affine invaraiant features. In *ICCV'07*, Oct. 14-20, 2007. 1, 2

[24] M. Yang, J. Yuan, and Y. Wu. Spatial selection for attentional visual tracking. In *CVPR'07*, June 17-22 2007. 1, 2

[25] Z. Yin and R. Collins. On-the-fly object modeling while tracking. In *CVPR'07*, June 17-22 2007. 1, 2

[26] T. Yu and Y. Wu. Differential tracking based on spatial-appearance model(SAM). In *CVPR'06*, volume 1, pages 720 – 727, 2006. 1, 2