Contextual Flow

Ying Wu and Jialue Fan
Northwestern University
2145 Sheridan Road, Evanston, IL 60208
{yingwu,jfa699}@eecs.northwestern.edu

Abstract

Matching based on local brightness is quite limited, because small changes on local appearance invalidate the constancy in brightness. The root of this limitation is its treatment regardless of the information from the spatial contexts. This papers leaps from brightness constancy to context constancy, and thus from optical flow to contextual flow. It presents a new approach that incorporates contexts to constrain motion estimation for target tracking. In this approach, one individual spatial context of a given pixel is represented by the posterior density of the associated feature class in its contextual domain. Each individual context gives a linear contextual flow constraint to the motion, so that the motion can be estimated in an over-determined contextual system. Based on this contextual flow model, this paper presents a new and powerful target tracking method that integrates the processes of salient contextual point selection, robust contextual matching, and dynamic context selection. Extensive experiment results show the effectiveness of the proposed approach.

1. Introduction

Matching points and regions is not only a fundamental component in motion analysis, but also a critical link in many applications such as target tracking, 3D recovery, and video analysis. A basic task is to estimate the motion between two consecutive frames. A common base for matching is on the local brightness and the constant brightness constraint (CBC) is often assumed. This leads to the well-known and elegant optical flow constraint and extensive research has been performed. However, CBC is quite limited and is often invalid in practice, simply because of small changes, such as image noise, illumination fluctuation and local deformation. Under such circumstances, the optical flow constraint does not always give the right constraint on the motion to facilitate robust target tracking.

The root of such a limitation is its treatment regardless of the information from rest of the image. As a matter of fact, any pixel is not isolated, but related to its spatial context that is induced by the pixels in its vicinity (near or far depending on the semantic levels). Although the brightness at one pixel may undergo large changes between two consecutive frames, its context shall be more stable and experience much less significant changes, because the context generally exhibits a constant pattern. For example, it is difficult to match the tip point of the nose. However, its context that gives a pattern of the nose or even the face shall make the matching less ambiguous. Therefore, integrating visual context into matching shall make it more accurate and more resilient to small local changes.

Visual context is a vague term, as it does not specify what the context is. Although context has not been well studied in motion analysis, there have been several ways in contextual modeling for object recognition. They can be roughly categorized based on the modeling of the spatial structure. *Structure-free* models ignore the spatial relations among pixels in the context, e.g., by using a bag of features or using feature histograms. The second category is *structure-flexible* contextual models that allow certain deformation of the spatial structure, e.g., the deformable templates, random graphs, and shape context. Another category is *structure-stiff* models that enforce strict spatial configurations or orderings, e.g., templates and spatial filters. In general, the enforcement of spatial structure tends to have a high precision in matching, but a low recall, because many actual good matches under small deformation may be missed. Structure-free models are generally easy to implement and have a higher recall, but they may give more false positives. Structure-flexible models compromise the two, but most of them tend to be complicated.

This paper presents a new approach that incorporates contexts to constrain motion estimation for target tracking. In the proposed contextual model, there are a number of feature classes, each of which is associated with an individual context. One individual context of a pixel is represented by the posterior density of this particular feature class at this pixel location, based on the densities of all features observed in its spatial contextual domain. Its total context is

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

the collection of these individual ones. For example, a pixel observes many other pixels with different edge directions in its vicinity, then its context on a particular edge direction (e.g., horizontal) is modelled by the posterior of having this particular edge direction at this pixel. As the density is less sensitive to small local appearance changes, matching this context leads to more robust and resilient results than matching brightness patterns.

The novelty of this paper includes the following three aspects. (1) It goes from matching local brightness to matching spatial contexts for motion analysis. (2) As a counterpart of the optical flow constraint, contextual flow constraints are derived and introduced in this paper. Corresponding to a set of feature contexts, these contextual flow constraints are linear constraints so that the motion can be estimated in an over-determined linear system. (3) Based on contextual matching, this paper presents a new and powerful target tracking method that integrates processes of salient context point selection, robust contextual matching, and dynamic context selection.

After a brief description on the related work in Sec. 2, the basic formulation of contextual matching and the contextual flow constraint are introduced in Sec. 3. Sec. 4 analyzes four important issues in this new approach, including salient contextual point determination, dynamic context selection, estimating rotation, and handling scales. Based on that, a new tracking method is presented in Sec. 5, and experiments are reported in Sec. 6.

2. Related Work

There have been extensive studies on optical flow in the literature. Optical flow can be determined locally (e.g., Lucas-Kanade's method [15]) or globally through regularization (e.g., Horn-Schunck's Method [12]). 3D motion can be recovered from the flow field based on flow models, or estimated directly through parametric flow and direct methods [5]. To alleviate the limitations of CBC, many approaches have been investigated, such as robust flow computation [4], dominant motion [16], and using subspace constraints [3, 10, 13]. We do not intend to list all here.

Beyond the use of raw brightness patterns, matching can be based on feature vectors. Local invariant features (such as SIFT [14]) can be obtained, and they can be used to match images of the same object from different views. Although they can be quite powerful in wide-baseline matching, it is not clear yet if they are appropriate for video tracking applications, because most of them are sensitive to local appearance changes such as local deformation. In addition, as these invariant descriptors are quite sophisticated, it is extremely difficult, if not impossible, to have differential forms, like optical flow, to enable computationally efficient gradient-based search that is important for target tracking. And thus the exhaustive search has to be performed.

Matching criteria that are suitable for target tracking have also been proposed, such as using SVM classification scores [1] and using color histograms. The mean-shift tracking method [7] and kernel-based tracking methods [6, 11, 8, 9] match color histograms between frames. Multiple kernel-based tracking methods have been studied for better tracking performance [11], for scale estimation [6], and for complex motions [8]. There methods are specific examples of matching color contexts. The concept of contextual flow introduced in this paper largely unifies these methods in a coherent treatment, reveals the connections to the optical flow methods, and provides further insights and new methods to motion analysis.

Modeling spatial context is a fundamental yet difficult issue in computer vision. At a low level, as a pixel is closely related to its neighborhood pixels, random fields have been exploited in modeling the joint distribution of the contexts. Because analytical solutions are only viable for low-order random fields, this limits their applications in motion analysis and target tracking. Another representation of spatial context uses filter responses that extract spatial features, which may reflect the higher-order relations to the context. A relation between filters and random fields is also studied [18]. The existing methods on the filter design are either pragmatic or biologically-motivated. At a higher-level, a pixel is closely related to the pixels that share the same semantic concepts. If the context can be captured, it may largely improve, for example, image segmentation [17] and shape recognition [2].

3. From Optical Flow to Contextual Flow

3.1. Constant Context Constraint

A pixel is located at $\mathbf{x} = [u, v]^T$, and its appearance is associated with a feature vector $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d$ in a feature space. We quantize the feature space to have a finite set of N discrete feature classes $\{\omega_1, \ldots, \omega_N\}$. In this paper, we refer a feature to a particular feature class for short. For example, a particular color (e.g., a pure red) or a particular type of texture is called a feature in this paper.

The set of features can be a mixture of various levels. They can be low-level such as color and texture features, and can also be at the object level (e.g. faces). In view of this, a pixel $\mathbf{x}$ can be associated with one or more features, i.e., we observe such features at the pixel location $\mathbf{x}$.

A pixel is not isolated, but is surrounded by its spatial contexts. For a pixel $\mathbf{x}$, its spatial contexts are constituted by the pixels in its *spatial context domain* $\Omega(\mathbf{x})$. Here we define $\Omega(\mathbf{x})$ as a circle centered at $\mathbf{x}$ with a radius r. Within $\Omega(\mathbf{x})$, an *individual context* of $\mathbf{x}$ consists of the pixels on a particular feature, i.e.,

$$\mathcal{C}_i = \{\mathbf{y} | \mathbf{f}(\mathbf{y}) \in \omega_i, \mathbf{y} \in \Omega(\mathbf{x})\},$$

and the *total context* of $\mathbf{x}$ is the union of all the individual contexts, i.e.,

$$\mathcal{C} = \bigcup_{i=1}^{N} \mathcal{C}_i.$$

To represent an individual context $\mathcal{C}_i$, we use its posterior density at $\mathbf{x}$, i.e., $p(\omega_i|\mathbf{x})$, that is estimated from $\mathcal{C}_i$ based on kernel density estimation. In this paper, we use 100 color contexts and 18 edge contexts. A color context is associated with quantized color, and an edge context with a quantized edge direction. For each feature ω_i, its *contextual map* $C_i(\mathbf{x})$ shows $p(\omega_i|\mathbf{x}, \Omega(\mathbf{x}))$ at a certain scale related to Ω. Fig. 1 shows examples of various contextual maps.

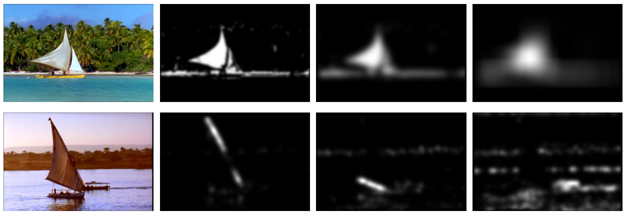

Figure 1. (Top) Examples of color contextual maps associated with the same color feature but at various scales. (Bottom) Examples of edge contextual maps associated with various edge features at the same scale.

It is clear that the constant brightness constraint (CBC), i.e., $I(\mathbf{x} + \Delta\mathbf{x}, t + \Delta t) = I(\mathbf{x}, t)$, is quite limited, as it rarely holds in practice. Here, we introduce a **constant context constraint** (CCC), where two pixels are matched if they have the same context:

$$p(\omega_i|\mathbf{x} + \Delta\mathbf{x}, t + \Delta t, \mathcal{C}) = p(\omega_i|\mathbf{x}, t, \mathcal{C}), \qquad (1)$$

where t is the time. We can omit $\mathcal{C}$ without confusion. This constraint says the motion $\Delta\mathbf{x}$ should not change the context and the context of the new location is the same as before the movement. Specifically, the posterior feature density of the point $\mathbf{x}$ keeps the same after the movement. This is more flexible than CBC, as it tolerates certain local deformation. If the features are carefully selected and designed, it is also robust to illumination fluctuations.

3.2. Contextual Flow Constraint Equations

Let's see how we go from optical flow to contextual flow. Based on the first order approximation for $p(\omega_i|\mathbf{x}+\Delta\mathbf{x}, t+\Delta t)$, it is clear that

$$\nabla_x^T p(\omega_i|\mathbf{x}, t)\Delta\mathbf{x} + \nabla_t p(\omega_i|\mathbf{x}, t)\Delta t = 0, \qquad (2)$$

where we call $\nabla_x p(\omega_i|\mathbf{x}, t)$ the *contextual gradient*, and $\nabla_t p(\omega_i|\mathbf{x}, t)$ the *contextual frame difference*. ∇_x denotes the derivatives w.r.t. $\mathbf{x}$, and ∇_t w.r.t. t.

We approximate $\nabla_t p(\omega_i|\mathbf{x}, t)\Delta t$ by the posterior density difference $p(\omega_i|\mathbf{x}, t+\Delta t) - p(\omega_i|\mathbf{x}, t)$ over the context domain $\Omega(\mathbf{x})$. It is clear that

$$p(\omega_i|\mathbf{x}, t) \propto p(\mathbf{x}|\omega_i, t)p(\omega_i|t)$$

where the prior $p(\omega_i|t)$ can also be easily estimated within a given context domain, and $p(\mathbf{x}|\omega_i, t)$ can be easily computed based on kernel density estimation,

$$p(\mathbf{x}|\omega_i, t) \propto \sum_{x_k \in \mathcal{C}_i} K(\mathbf{x}, \mathbf{x}_k|t)$$

where $K(,\cdot,)$ is a kernel function.

The following shows that the contextual gradient $\nabla_x p(\omega_i|\mathbf{x}, t)$ can be computed very easily. We define a local *conditional shift* $\mu_i(\mathbf{x})$ vector for each context $\mathcal{C}_i$ at time t (here we omit t without confusion):

$$\mu_i(\mathbf{x}) \triangleq E\{(\mathbf{y}-\mathbf{x})|\mathbf{y} \in \omega_i\} = \frac{1}{Z_i(\mathbf{x})}\int_\Omega (\mathbf{y}-\mathbf{x})p(\mathbf{y}|\omega_i)d\mathbf{y},$$

where $Z_i(\mathbf{x}) = \int_\Omega p(\mathbf{y}|\omega_i)d\mathbf{y} \simeq V_\Omega p(\mathbf{x}|\omega_i)$. We approximate the conditional $p(\mathbf{y}|\omega_i)$ by its Taylor series at $\mathbf{x}$:

$$p(\mathbf{x}|\omega_i)+\nabla_x^T p(\mathbf{x}|\omega_i)(\mathbf{y}-\mathbf{x})+\frac{1}{2}(\mathbf{y}-\mathbf{x})^T\nabla_x^2 p(\mathbf{x}|\omega_i)(\mathbf{y}-\mathbf{x}).$$

Noticing Ω is symmetric, we have

$$\mu_i(\mathbf{x}) = c\frac{\nabla_x p(\mathbf{x}|\omega_i)}{p(\mathbf{x}|\omega_i)}, \quad \text{where } c = \frac{r^2}{2}$$

Similarly, we can define a local *total shift*:

$$\mu_0(\mathbf{x}) \triangleq E\{(\mathbf{y}-\mathbf{x})|\mathbf{y} \in \Omega\} = c\frac{\nabla_x p(\mathbf{x})}{p(\mathbf{x})},$$

which is the average of all the conditional shifts for various contexts. To compute the contextual gradient, we have,

$$\begin{aligned}\nabla_x p(\omega_i|\mathbf{x}) &= \nabla_x \left\{p(\omega_i)\frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x})}\right\} \\ &= \frac{1}{c}p(\omega_i|\mathbf{x})\Big[\mu_i(\mathbf{x}) - \mu_0(\mathbf{x})\Big] \qquad (3)\end{aligned}$$

Putting Eq. 3 into Eq. 2, for each context $\mathcal{C}_i$, we have a **contextual flow constraint equation**:

$$\Big[\mu_i(\mathbf{x}) - \mu_0(\mathbf{x})\Big]^T\Delta\mathbf{x} + c\left[\frac{p(\omega_i|\mathbf{x}, t+1)}{p(\omega_i|\mathbf{x}, t)} - 1\right] = 0, \quad (4)$$

where $\Delta\mathbf{x}$ is the *contextual flow* and we use $\Delta t = 1$ without loss of generality. When we denote

$$\bar{\mu}_i(\mathbf{x}) = \mu_i(\mathbf{x}) - \mu_0(\mathbf{x}), \text{ and } b_i = c\left[1 - \frac{p(\omega_i|\mathbf{x}, t+1)}{p(\omega_i|\mathbf{x}, t)}\right],$$

we see a simple linear constraint on motion for context $\mathcal{C}_i$:

$$\boxed{\bar{\mu}_i(\mathbf{x})^T \Delta\mathbf{x} - b_i = 0.} \tag{5}$$

In practice, the conditional shift $\mu_i(\mathbf{x})$ and the total shift $\mu_0(\mathbf{x})$ can be easily estimated by:

$$\hat{\mu}_i(\mathbf{x}) = \frac{1}{n_i}\sum_{y\in\mathcal{C}_i}(\mathbf{y}-\mathbf{x}) = \left(\frac{1}{n_i}\sum_{y\in\mathcal{C}_i}\mathbf{y}\right) - \mathbf{x}$$

where n_i is the number of samples in $\mathcal{C}_i$, and

$$\hat{\mu}_0(\mathbf{x}) = \frac{1}{n}\sum_{y\in\mathcal{C}}(\mathbf{y}-\mathbf{x}),$$

where n is the total number of samples in Ω.

It is clear from Eq. 4 that each individual context contributes a linear constraint on the motion. Each constraint is weighted by $W_i(\mathbf{x}) = p(\omega_i|\mathbf{x}, t)$, meaning a stronger context has a larger weight. We denote by $\mathbf{W}(\mathbf{x}) \triangleq diag[W_1(\mathbf{x}), \dots, W_N(\mathbf{x})]$. Considering all $N > 2$ contexts, the total context gives a weighted over-determined linear system, called *contextual system*. We denote by

$$\mathbf{U}_r(\mathbf{x}) \triangleq \left[\bar{\mu}_1(\mathbf{x}), \dots, \bar{\mu}_N(\mathbf{x})\right]^T, \quad \mathbf{b}_r(\mathbf{x}) \triangleq [b_1, b_2 \dots, b_N]^T,$$

$$\mathbf{U}(\mathbf{x}) = \mathbf{W}(\mathbf{x})\mathbf{U}_r(\mathbf{x}), \qquad \mathbf{b}(\mathbf{x}) = \mathbf{W}(\mathbf{x})\mathbf{b}_r(\mathbf{x})$$

and we have

$$\mathbf{U}(\mathbf{x})\Delta\mathbf{x} = \mathbf{b}(\mathbf{x}), \text{ or simply } \mathbf{U}\Delta\mathbf{x} = \mathbf{b}. \tag{6}$$

If $\mathtt{rank}(\mathbf{U}) = 2$, the weighted least squares solution, $\Delta\mathbf{x} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{b}$, of this overdetermined system leads to a unique determination of the contextual flow at $\mathbf{x}$, and thus the motion at $\mathbf{x}$. In addition, we can also adjust the weights adaptively (see Sec. 4.2) based on the consistency of individual contexts.

This new method is conceptually different from mean-shift tracking, although they both use kernel density estimation. Mean-shift method matches histograms and explicitly gives a motion vector that maximizes the density of the re-weighted pixels, while our method matches the contextual posteriors and gives a set of linear constraints on motion that enables easy dynamic context selection and adaptation.

3.3. Two Extensions

It is clear from the above that the contextual flow $\Delta\mathbf{x}$ can be determined by the *contextual system* associated with $\mathbf{x}$ in its basic form in Eq. 6, i.f.f. the system is of a full rank, i.e., $\mathtt{rank}(\mathbf{U}) = 2$ or $\Delta\mathbf{x}$ is fully observable from its evidence (or observations) $\mathbf{b}$. There exist cases where $\mathbf{U}$ is rank deficient, and thus the solution is not unique.

This difficulty of rank deficiency can be alleviated by considering the pixels in the vicinity of $\mathbf{x}$. We assume a set of pixel locations $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, each of which is associated with a set of context flow constraints, i.e.,

$$\mathbf{U}_i(\mathbf{x}_i)\Delta\mathbf{x}_i = \mathbf{b}(\mathbf{x}_i), \text{ or simply } \mathbf{U}_i\Delta\mathbf{x}_i = \mathbf{b}_i$$

where $\Delta\mathbf{x}_i$ is the motion for pixel $\mathbf{x}_i$.

If they share the same motion, i.e., $\Delta\mathbf{x}_i = \Delta\mathbf{x}$, we have an extended Lucas-Kanade method. We have

$$\begin{bmatrix}\mathbf{U}_1\\ \dots \\ \mathbf{U}_m\end{bmatrix}\Delta\mathbf{x} \triangleq \mathbf{U}_c\Delta\mathbf{x} = \begin{bmatrix}\mathbf{b}_1\\ \dots \\ \mathbf{b}_m\end{bmatrix} \tag{7}$$

where $\mathbf{U}_c$ is a direct concatenation of $\{\mathbf{U}_i\}$. We can also assign weights to different $\mathbf{x}_i$. It is clear that $\mathbf{U}_c$ is more likely to have a full rank than any individual $\mathbf{U}_i$.

If their motion are correlated and are constrained in a subspace, we have

$$\begin{bmatrix}\Delta\mathbf{x}_1\\ \dots \\ \Delta\mathbf{x}_m\end{bmatrix} = \mathbf{Bm}$$

where $\mathbf{B}$ is constituted by a set of basis vectors, and $\mathbf{m}$ is the component motion. For example, the set of pixels share an affine motion, and we can write:

$$\begin{bmatrix}\Delta\mathbf{x}_1\\ \Delta\mathbf{x}_2\\ \dots \\ \Delta\mathbf{x}_m\end{bmatrix} = \begin{bmatrix} u_1 & v_1 & 0 & 0 & 1 & 0\\ 0 & 0 & u_1 & v_1 & 0 & 1\\ u_2 & v_2 & 0 & 0 & 1 & 0\\ 0 & 0 & u_2 & v_2 & 0 & 1\\ \dots & \dots & \dots & \dots & \dots & \dots\\ u_m & v_m & 0 & 0 & 1 & 0\\ 0 & 0 & u_m & v_m & 0 & 1\end{bmatrix}\begin{bmatrix}m_1\\ m_2\\ m_3\\ m_4\\ m_5\\ m_6\end{bmatrix}$$

Of course, the motion subspace can also be learned if training examples are available.

Putting the subspace motion to the context flow constraints, we have a generalized subspace contextual flow:

$$\begin{bmatrix}\mathbf{U}_1 & & & \\ & \mathbf{U}_2 & & \\ & & \dots & \\ & & & \mathbf{U}_m\end{bmatrix}\mathbf{Bm} \triangleq \mathbf{U}_s\mathbf{Bm} = \begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\\ \dots \\ \mathbf{b}_m\end{bmatrix} \tag{8}$$

where $\mathbf{U}_s$ is a block diagonal concatenation of $\{\mathbf{U}_i\}$.

4. Beyond Basic Contextual Flow

4.1. Salient Contextual Point Determination

For any pixel location $\mathbf{x}$, its contextual flow is associated with its contextual system, which is characterized by $\mathbf{U}(\mathbf{x})$ in its basic form, and $\mathbf{U}_c(\mathbf{x})$ and $\mathbf{U}_s(\mathbf{x})$ in its extended forms. The quality of the solution to the contextual flow is largely determined by the characteristics of the contextual system.

If the contextual system is close to singular, meaning that a small perturbation of the system (i.e., $\mathbf{U} + \Delta\mathbf{U}$) can change the solution to $\Delta\mathbf{x}$ dramatically. Such small perturbations may be induced by many factors such as image noises, feature quantization, and small changes in visual appearances, which are inevitable and common in practice. For example, if the context domain of $\mathbf{x}$ exhibits small deformation, or is influenced by small amount of occlusion, there will be small changes in $\mathbf{U}$. For a close-to-singular contextual system, such small changes will lead to drastic changes in estimation of the contextual flow, which is certainly undesirable.

Therefore, it is of great interest to determine the pixel locations that can give robust contextual flow. We call such points *salient contextual points*. It is clear that the contextual systems associated with such locations have to be far from singular. Here we use the condition number $\kappa(\mathbf{x})$ of $\mathbf{U}(\mathbf{x})$, which is the ratio between the largest singular value and the smallest singular value of $\mathbf{U}(\mathbf{x})$. It is clear that $\kappa(\mathbf{x}) \geq 1$.

In practice, we can have two ways to determine salient contextual points. A simple method is to threshold $\kappa(\mathbf{x})$, i.e., $\mathbf{x}$ is a salient point if $\kappa(\mathbf{x}) \geq \kappa_0$. Without using a threshold, another way is to find the local minima of $\kappa(\mathbf{x})$. Some examples are shown in Fig. 2 by using edge contexts.

As such salient contextual points are more resilient to many local perturbations, matching them in different image frames (at different time instants or from different views) is more reliable. Thus, they tend to have better repeatability in matching and tracking. To tracking a target, we use a set of salient contextual points when applying the extended Lucas-Kanade method and the subspace context flow, as described in section 3.3.

4.2. Dynamic Context Selection

In the basic form of contextual flow determination, we assume various contexts are equally reliable. As different contexts exhibit different reliabilities over time, we assign additional weights to the individual contexts, and the weighted least squares solution actually minimizes the following objective function:

$$E(\Delta\mathbf{x}) = \sum_{i=1}^{N} \alpha_i W_i(\mathbf{x}) ||\bar{\mu}_i(\mathbf{x})^T \Delta\mathbf{x} - b_i||^2, \qquad (9)$$

where α_i is the *reliability weight* for context ω_i.

In target tracking, we collect statistics so as to adapt these reliability weights over time. We call this process *dynamic context selection*. For each ω_i, once we have estimated the contextual flow $\hat{\Delta}\mathbf{x}$, we compute its fitting error at every frame, i.e., we have $e_i(t) = ||\bar{\mu}_i(\mathbf{x})^T \hat{\Delta}\mathbf{x} - b_i||^2$. In a sliding time window, we can compute the variance $\sigma_i(t)$ based on each $e_i(t)$. $\sigma_i(t)$ reflects the uncertainty of context ω_i and its reliability. A larger σ_i implies a less reliable context. Thus, we use its inverse to weight the context, i.e.,

$$\alpha_i(t) = \frac{1}{\sigma_i(t)} \Big/ \Big(\sum_{i=1}^{N} \frac{1}{\sigma_i(t)} \Big),$$

which reduces the influence of the unreliable contexts.

In addition to the above adaptation process over time, robust statistics techniques can be also applied at a certain time instant. There are cases when some contexts become outliers, i.e., they provide completely different estimates from others. By examining the fitting errors $\{e_i\}_{i=1}^{N}$, apparent outliers can be spotted. More sophisticated methods such as RANSAC can also be applied here.

In its extended forms, as we have a number of m anchor points and a number of N individual contexts, the objective in determining the contextual flow is:

$$E(\Delta\mathbf{x}) = \sum_{j=1}^{m} \beta_j \sum_{i=1}^{N} \alpha_i W_i(\mathbf{x}) ||\bar{\mu}_i(\mathbf{x}_j)^T \Delta\mathbf{x} - b_{ij}||^2, \quad (10)$$

where α_i weights the reliability of context i and β_j for anchor point $\mathbf{x}_j$. The same principle of adaptation and outlier detection can also be applied to the anchor points. This is especially useful to handle partial occlusion.

4.3. Estimating Rotation

Some features, such as color, are rotation invariant, and thus the corresponding contexts are rotation invariant. On one hand, it is good as matching such contexts is insensitive to rotation. On the other hand, we cannot estimate rotation based on such contexts, as rotation is unobservable.

Some other features, such as image gradients and edge directions, are not invariant to rotation. As image gradient provides clues to the the shape of the target, it provides more accurate matching and alignment than rough color features. However, its matching has to optimize over rotation besides the location. Resolving this extra degree of freedom demands more computation in estimating rotation.

Here we introduce a new method that is able to provide computationally efficient rotation-invariant matching and is able to estimate rotation at the same time. The features we use here are the directions of image gradient. We quantize them to a number of N directions (e.g., $N = 18$). Each individual context is constituted by the points having the same edge direction, and the points are weighted by their image gradient magnitudes, i.e.,

$$\mathcal{C}_i = \{(\mathbf{y}, \gamma) | \angle \nabla I(\mathbf{y}) = \theta_i, \mathbf{y} \in \Omega\},$$

where $\nabla I(\mathbf{y})$ is the image gradient at $\mathbf{y}$, θ_i is the angle of the edge direction ($\theta \in [0, \pi)$), and $\gamma = |\nabla I(\mathbf{y})|$ is the weight of $\mathbf{y}$.

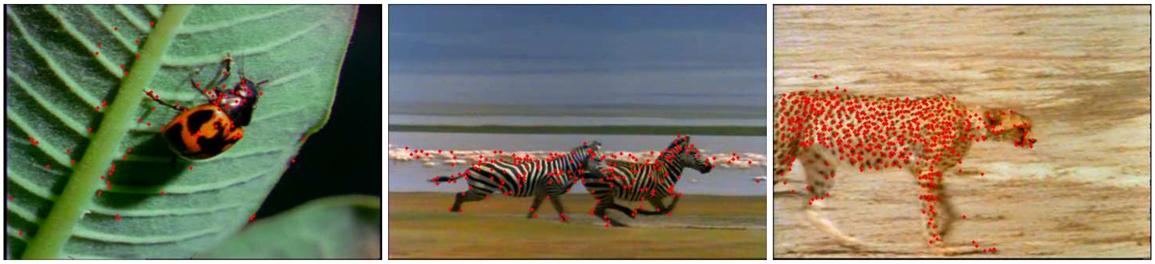

Figure 2. Salient anchor point determination. Red dots show the selected salient contextual points.

It is clear that the context consistency constraints in Eq. 1 and Eq. 2 are not satisfied if there is a rotation. However, the constraint actually holds up to a shift, i.e.,

$$p(\omega_i|\mathbf{x}+\Delta\mathbf{x}, t+\Delta t, \mathcal{C}) = p(\omega_j|\mathbf{x}, t, \mathcal{C}), \tag{11}$$

where $j = mod(i+k, N)$, and k is a constant but unknown shift (i.e., context ω_i matches another one). Considering all the contexts in the contextual system, we have:

$$\mathbf{U}\Delta\mathbf{x} = \mathbf{S}\mathbf{b} \tag{12}$$

where $\mathbf{S}$ is a circular shift matrix of an unknown shift. For example, if $k = 2$ and $N = 5$,

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{I}_3 \\ \mathbf{I}_2 & 0 \end{bmatrix}$$

Both $\Delta\mathbf{x}$ and $\mathbf{S}$ are unknowns in Eq. 12. A conventional approach is to estimate them in an EM framework, but it may not give the global optimum. Considering the special characteristics of the shift matrix, our new solution is to match them in the frequency domain. The unknown shift does not change the magnitude of the Fourier coefficients, i.e.,

$$|\mathbf{F}\mathbf{U}\Delta\mathbf{x}| = |\mathbf{F}\mathbf{b}|, \tag{13}$$

where $\mathbf{F}$ is the DFT matrix of the discrete Fourier transformation. Eq. 13 does not have the unknown shift $\mathbf{S}$ anymore. In practice, we only need to use a subset of low-frequency Fourier coefficients in matching. Once $\Delta\mathbf{x}$ is determined from Eq. 13, rotation can be solved straightforwardly by checking the shift, if needed.

4.4. Handling Scale

It is in general difficult to handle scale in matching, as most known simple features are not scale invariant. Here we give a new method in handling scale following the same idea as in handling rotation.

A context is not only determined by its associated feature class, but also by its effective context domain that is determined by the scale. For a certain context class ω_i, we collect a series of contexts over a spectrum of effective context domains, i.e.,

$$\mathcal{C}_i^k = \{\mathbf{y}|\mathbf{f}(\mathbf{y}) \in \omega_i, \mathbf{y} \in \Omega_k(\mathbf{x})\},$$

where Ω_k is the effective context domain at scale s_k. The context consistency constraint still holds up to a scale shift:

$$p(\omega_i^k|\mathbf{x}+\Delta\mathbf{x}, t+\Delta t, \mathcal{C}) = p(\omega_i^{k+\tau}|\mathbf{x}, t, \mathcal{C}), \tag{14}$$

where τ is constant, unknown but small shift in the scale spectrum. For a certain type of context ω_i, once we collect all the constraints over a scale spectrum, we have

$$\mathbf{U}_i\Delta\mathbf{x} = \mathbf{S}\mathbf{b}_i \tag{15}$$

where $\mathbf{S}$ is a contaminated shift matrix. For example, if the target is zoomed in by τ levels in the scale spectrum:

$$\mathbf{S} = \begin{bmatrix} 0 & \mathbf{I}_{K-\tau} \\ \mathbf{N}_\tau & 0 \end{bmatrix}$$

where K is the range of scale spectrum, and $\mathbf{N}$ is a random matrix. In practice, τ is generally small because we generally do not have dramatic zooming in and out in video. So we can still do the matching in the Fourier domain, i.e.,

$$|\mathbf{F}\mathbf{U}_i\Delta\mathbf{x}| = |\mathbf{F}\mathbf{b}_i| \tag{16}$$

If we consider all classes of contexts, we minimize the following objective function for motion estimation:

$$E(\Delta\mathbf{x}) = \sum_{i=1}^{N} (|\mathbf{F}\mathbf{U}_i\Delta\mathbf{x}| - |\mathbf{F}\mathbf{b}_i|)^2$$

5. Contextual Target Tracking

Based on the above contextual flow techniques, we design a new target tracking method. Shown in Fig. 4, it has three major components:

- *Salient contextual points selection.* Once the target region in initialized, we can select a number of m contextual anchor points according to their saliency using the method described in sec. 4.1. Each anchor point contributes N constraints on the target motion. This is an early selection process.

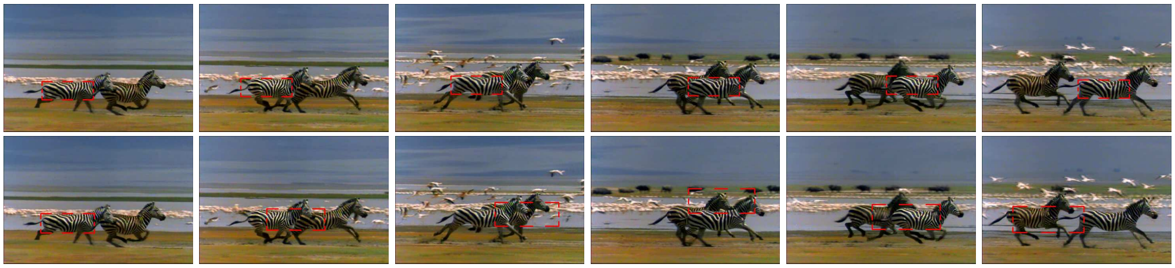

Figure 3. A comparison of contextual flow tracking and mean-shift. (top) the proposed method, and (bottom) Mean-shift tracker.

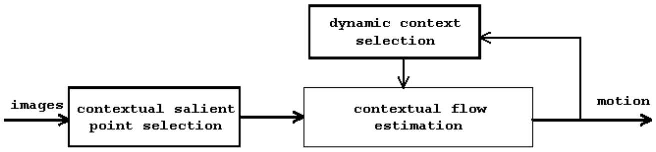
dynamic context selection
images
contextual salient point selection
contextual flow estimation
motion

Figure 4. The diagram of context target tracking.

- *Contextual flow determination.* Based on the constraints from m anchor points, we compute the contextual flow using the extended Lucas-Kanade method 7 or the generalized subspace contextual flow method 8. If rotation and scaling are needed, we use the methods described in Sec. 4.3 and Sec. 4.4. As contextual flow reflects the velocity field, a line search is needed to determine the actual displacement and motion, as in the treatment in the optical flow-based matching methods.
- *Dynamic context learning and selection.* We keep a sliding time window of T frames for adaptation. After collecting the matching variances for all the contexts, we re-weight these contexts. In addition, we also re-weight the importance of the anchor points. This is a late and on-line selection process.

This algorithm is implemented in C++ and tested on Pentium-IV 3Ghz desktop. Without code optimization, the program runs comfortably at $15 - 20$fps on average.

6. Experiments

6.1. Handling Local Appearance Changes

Tracking targets undergoing local deformation is difficult in practice. As contextual matching is insensitive to local deformation, it handles quite well the appearance changes due to local deformation in our experiments. We use 18 edge contexts in our CFT and 10∼25 contextual anchor points in the extended Lucas-Kanade method. A comparison example between CFT and a Mean-Shift tracker that is implemented in an enhanced YCbCr space with 1040 bins is given in Fig. 3. Several other examples of CFT are shown in Fig. 6. The scale of Mean-shift tracker becomes unstable when nearby background exhibits similar color histograms, while the CFT is quite robust to the local deformation due to the constancy of the context.

We manually labelled the ground truth of our testing sequences for evaluation. Fig. 5 shows the comparison between CFT and Mean-Shift in tracking error over time on the `zebra` sequence. It clear that CFT outperforms.

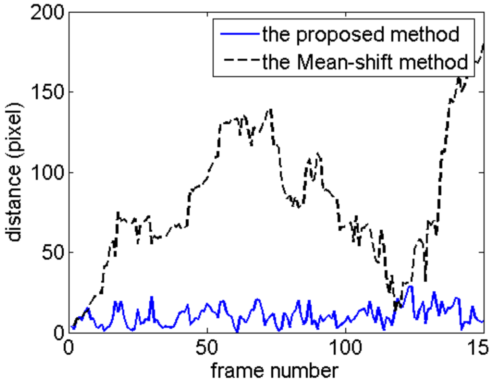

Figure 5. The comparison of tracking errors between contextual tracking and Mean-Shift.

We show an example of CFT in handling partial occlusion in Fig. 6(3rd row), where the target undergoes severe occlusion when passing the tree and the sign. The context plays a critical role in finding the right matches. More examples can be viewed in the supplementary materials.

6.2. Handling rotation and Scaling

It is in general difficult to handle rotation and scaling in differential tracking methods. These two issues are especially difficult for the Mean-Shift tracker, because rotation is unobservable in matching color histograms and a single bandwidth is used in estimate the histogram in Mean-Shift. These two issues can be handled in contextual flow tracking by using edge contexts and multiple context domains. Two examples are shown in Fig. 7, and more can be seen in the supplementary materials.

7. Conclusion

This paper presents a new concept of contextual flow, based on which motion analysis can be performed on contextual matching. As an image pixel is by no means isolated but closely tied to its spatial context, matching its contexts between image frames is more resilient to local appearance changes than matching its brightness. A certain feature, which can be at any semantic level, induces one individual

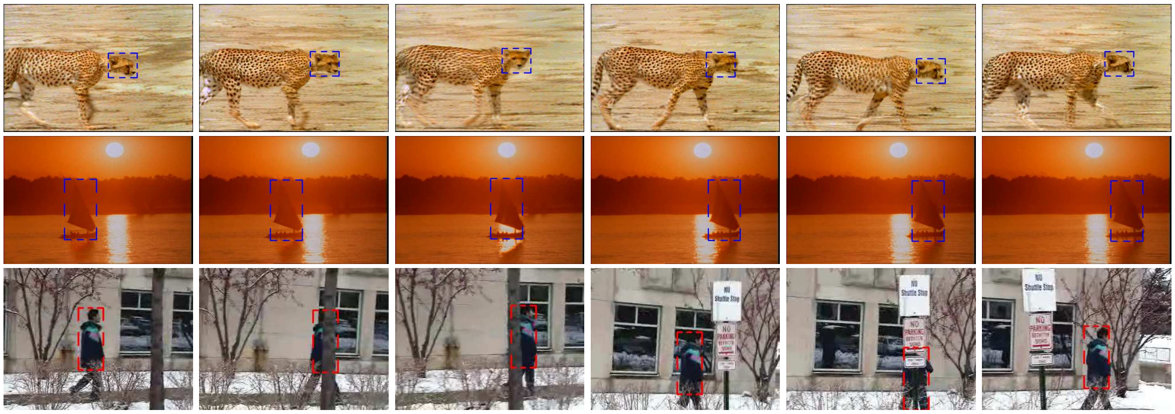

Figure 6. Examples of contextual tracking in handling local appearance changes due to local deformation, lighting and occlusion.

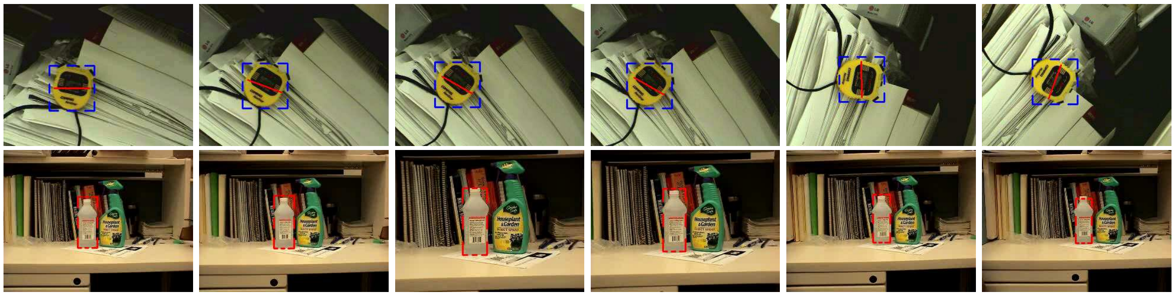

Figure 7. Estimating rotation and scaling. (top) The red line shows the orientation of the target estimated by contextual matching. (bottom) Contextual flow tracking is able to estimate scale quite well because of the matching of contexts at different scales.

context that is represented by the belief of having this feature class at the pixel of interest. Each individual context in turn contributes one linear contextual flow constraint on the motion at this pixel location. Motion is determined by all the constraints from various contexts. Based on the contextual flow model, a new target tracking method is designed and is integrated with the salient contextual point selection and the dynamic context selection processes.

Acknowledgement

This work was supported in part by National Science Foundation grant IIS-0347877 and US Army Research Laboratory and the US Army Research Office under grant ARO W911NF-08-1-0504.

References

[1] S. Avidan. Support vector tracking. *IEEE T-PAMI*, 26:1064–1072, Aug. 2004. 2

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE T-PAMI*, 24:509–522, 2002. 2

[3] J. Bergen, P. Burt, R. Hingorani, and S. Peleg. A three-frame algorithm for estimating two-component image motion. *IEEE T-PAMI*, 14, Sept. 1992. 2

[4] M. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63:75–104, 1996. 2

[5] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *IEEE CVPR*, pages 8–15, Santa Barbara, CA, June 1998. 2

[6] R. Collins. Mean-shift blob tracking through scale space. In *IEEE CVPR*, Madison, WI, Jun. 2003. 2

[7] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE CVPR*, volume II, pages 142–149, Hilton Head Island, South Carolina, 2000. 2

[8] Z. Fan, M. Yang, and Y. Wu. Multiple collaborative kernel tracking. *IEEE T-PAMI*, 29:1268–1273, July 2007. 2

[9] Z. Fan, M. Yang, Y. Wu, G. Hua, and T. Yu. Efficient optimal kernel placement for reliable visual tracking. In *IEEE CVPR*, New York City, June 2006. 2

[10] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE T-PAMI*, 20:1025–1039, 1998. 2

[11] G. Hager, M. Dewan, and C. Stewart. Multiple kernel tracking with SSD. In *IEEE CVPR*, Washington, DC, Jun. 2004. 2

[12] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981. 2

[13] M. Irani. Multi-frame optical flow estimation using subspace constraints. In *IEEE ICCV*, Corfu, Greece, Sept. 1999. 2

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. 2

[15] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Int'l Joint Conf. on Artificial Intelligence*, pages 674–679, Vancouver, Aug. 1981. 2

[16] H. Sawhney and S. Ayer. Compact representations of videos through dominant and multiple motion estimation. *IEEE T-PAMI*, 18:814–830, 1996. 2

[17] Z. Tu. Auto-context and its application to high-level vision tasks. In *IEEE CVPR*, Anchorage, Alaska, June 2008. 2

[18] S. C. Zhu, Y. N. Wu, and D. B. Mumford. FRAME: Filters, random field and maximum entropy: — towards a unified theory for texture modeling. *Int'l J. on Computer Vision*, 27:1–20, 1998. 2