Aligning Spatio-Temporal Signals on a Special Manifold

Ruonan Li and Rama Chellappa

Center for Automation Research, University of Maryland
College Park,MD, 20742, USA
`{liruonan,rama}@umiacs.umd.edu`

Abstract. We investigate the spatio-temporal alignment of videos or features/signals extracted from them. Specifically, we formally define an *alignment manifold* and formulate the alignment problem as an optimization procedure on this non-linear space by exploiting its intrinsic geometry. We focus our attention on semantically meaningful videos or signals, *e.g.*, those describing or capturing human motion or activities, and propose a new formalism for temporal alignment accounting for executing rate variations among realizations of the same video event. By construction, we address this static and deterministic alignment task in a dynamic and stochastic manner: we regard the search for optimal alignment parameters as a recursive state estimation problem for a particular dynamic system evolving on the alignment manifold. Consequently, a Sequential Importance Sampling iteration on the alignment manifold is designed for effective and efficient alignment. We demonstrate the performance on several types of input data that arise in vision problems.

1 Introduction

In this paper, we consider the problem of aligning two spatio-temporal signals (*i.e.*, videos, their filtered versions, or spatio-temporal features extracted from them.) which come from the same dynamic scene or the same category of dynamics. The misalignment between the two signals, captured by distinct cameras at the same time or by the same camera at different times, may result from the differences in view points, view angles, internal calibration parameters, as well as temporal shifts and scaling. Previous work on video sequence alignment mostly used feature-based approaches [1–7] or direct approaches [8–10]. In the former class, features like two-frame correspondences of interest points or trajectories of tracked objects were used as inputs to the alignment algorithm, while in the latter, intensity, color, or other pixel/patch level appearance attributes were used. The spatial aspect of the misalignment was mostly modeled as one of the transforms including affine, homography, and perspective ones between the image plane coordinates of the two signals, based on different assumptions made regarding imaging conditions. The temporal misalignment, on the other hand, mainly took frame rate and shift synchronization into account, modeled as a 1-D affine transform along the time axis. The algorithms were designed

for parametric representations of the particular transforms to achieve optimal alignments. The warping parameters were then obtained using a numerical optimization method which is typically an exhaustive search or a greedy method such as gradient descent.

The first step taken in this work is to revisit the issue of temporal misalignment, which comes not only from the camera aspect (frame-rate and temporal shift), but also from the observed dynamics. We look into semantically meaningful visual dynamics beyond plain spatio-temporal volumes: one of the examples of semantically meaningful visual signals is videos recording human actions/activities. The same class of activities (*e.g.*, walking) may contain realizations executed at varying rates, though the essential characterization for that activity category is rate independent. This rate change is in fact a temporal misalignments among realizations (signals) and is described by a non-affine time warping [11, 12]. Therefore, a complete description of the temporal misalignment regarding these signals should include time warping as well. A second concern is about the spatial aspect of the alignment algorithm, which usually pertains particularly to either feature-based methods or direct methods and sticks to the parametric spatial transform assumed. Existing algorithms are far from being scalable and flexible to easily adapt to different parametric model and different inputs. Moreover, it is always crucial to strike a balance between computational complexity and convergence towards global optimum.

Taking all these factors into account, we reformulate the spatio-temporal alignment problem and provide a general framework and associated computational algorithms. Specifically, we propose the concept of the *alignment manifold*, which is the nonlinear space of all possible spatio-temporal transformations with an intrinsic geometric characterization. We detail the construction of the alignment manifold and discuss basic manipulations of the elements on it. The spatio-temporal signal alignment, consequently, becomes an optimization procedure on the manifold, regardless of whether the inputs are features or appearances, provided that an objective function is properly defined to measure the misalignment of the two signal under a spatio-temporal transformation model. In particular, we present a Bayesian optimization algorithm on the manifold based on Sequential Importance Sampling (SIS) [13] , to achieve both efficiency and better convergence to the global optimum. The key idea is to regard the optimal alignment as a static state to be recursively estimated from the observed misalignment such that the posterior probability density of the estimated state reaches maximum at the true optimal alignment.

In short, the contributions of this paper are (1) we present a general framework for spatio-temporal alignment, incorporating temporal warping and various parametric spatial transforms as well as inputs; (2) we introduce the *alignment manifold*, a manifold tuned to the alignment task; and (3) a SIS algorithm is specifically designed for the alignment manifold to generate the numerical solution to the alignment problem.

2 The Framework of Alignment Problem

Given 3-dimensional spatio-temporal signals S^1 and S^2, whose elements are denoted as $S^1(x, y, t)$ and $S^2(x, y, t)$ respectively, the spatio-temporal alignment problem aims to solve the following optimization problem

$$\min_{\mathbf{p} \in \mathfrak{M}} J(S^1, S^2, \mathbf{p}) \tag{1}$$

where $\mathbf{p}$ is the parameter vector specifying the alignment transform, $\mathfrak{M}$ is the alignment manifold, $i.e.$, the space of all feasible $\mathbf{p}$'s, and J is a measure of misalignment to be minimized by an optimal $\mathbf{p}$. As in previous efforts, we assume the relative internal and external parameters of the two cameras to be fixed but unknown, $i.e.$, both stationary or jointly moving. As a result, the spatial misalignment and temporal misalignment become decoupled. In other words, we may split $\mathbf{p}$ into two components as $\mathbf{p} = [\mathbf{p}_S^T, \mathbf{p}_T^T]^T$, so that the spatial and temporal misalignment can be independently handled. (Cameras with relative motion and coupled spatio-temporal misalignment are important situations though beyond the scope of this work.) The alignment manifold $\mathfrak{M}$ is accordingly decomposed into the Cartesian product of two submanifolds as $\mathfrak{M} = \mathfrak{M}_S \times \mathfrak{M}_T$, where $\mathbf{p}_S \in \mathfrak{M}_S$ and $\mathbf{p}_T \in \mathfrak{M}_T$. The explicit analytical form of J depends on the specific spatial and temporal transform involved, as well as the measure of misalignment. We give three examples for illustrative purposes.

Example 1 S^1 and S^2 are grey-level videos, the spatial displacement is 2-D affine, and temporal transform is 1-D affine. The misalignment is measured as the pixel-wise mean square error. In this case, $J(S^1, S^2, \mathbf{p}) = \sum_{x,y,t}(S^1(x, y, t) - S^2(x + u, y + v, t + w))^2$, and

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & 0 & b_1 \\ a_{21} & a_{22} & 0 & b_2 \\ 0 & 0 & a & b \end{bmatrix} \begin{bmatrix} x \\ y \\ t \\ 1 \end{bmatrix}. \tag{2}$$

The corresponding alignment parameter vectors are $\mathbf{p}_S = (a_{11}, a_{12}, a_{21}, a_{22}, b_1, b_2)^T$ and $\mathbf{p}_T = (a, b)^T$ with $\mathfrak{M}_S$ to be the 2-D affine group $\mathbb{A}(2)$ and $\mathfrak{M}_T$ to be $\mathbb{R}^+ \times \mathbb{R}$.

Example 2 S^1 and S^2 are color videos, $i.e.$, S^i contain three channels S_j^i, $j = 1, 2, 3$, spatial transform is 2-D homography, and the temporal transform is a nonlinear time warping. The misalignment is measured as the pixel-wise mean square error of the intensity. In this case, $J(S^1, S^2, \mathbf{p}) = \sum_j \alpha_j \sum_{x,y,t}(S_j^1(x, y, t) - S_j^2(x', y', t'))^2$, $x' = \frac{h_{11}x + h_{12}y + h_{13}}{h_{31}x + h_{32}y + h_{33}}$, $y' = \frac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + h_{33}}$, and $t' = W(t)$, where α_j's are the weights for the channels and $W(t)$ is the time warping function. If we denote $H = [h_{i,j}]_{3 \times 3}$ to be the homography matrix with the constraint of unit determinant ($i.e.$ $\det H = 1$, without loss of generality), then we have $\mathbf{p}_S = H$, $\mathbf{p}_T = W$, $\mathfrak{M}_S$ is the 3×3 special linear group $\mathbb{SL}(3)$, and $\mathfrak{M}_T$ is the set of all possible time warpings.

Example 3 S^1 and S^2 contain N spaces-time point trajectories respectively, $i.e.$, $S^i = \{T_j^i\}_{j=1,2,\cdots,N}$ and $T_j^i = \{(x_j^i(t), y_j^i(t))\}_t$, where $(x_j^1(t), y_j^1(t))$

$(x_j^2(t'), y_j^2(t'))$ are assumed to come from the jth tracked interest point corresponding to the same 3-D point, captured by two pinhole cameras. Then considering perspective misalignment of the trajectories we have $J(S^1, S^2, \mathbf{p}) = \sum_j \sum_t \|[x_j^1(t), y_j^1(t), 1]\mathbf{F}[x_j^2(W(t)), y_j^2(W(t)), 1]^T\|^2$. Here $\mathbf{F}$ is the 3×3 fundamental matrix, and we may regard $\mathbf{p}_S = \mathbf{F}$ and $\mathfrak{M}_S$ to be the set of all possible fundamental matrices.

3   The Alignment Manifold

In this section we look into the alignment manifold $\mathfrak{M} = \mathfrak{M}_S \times \mathfrak{M}_T$, whose elements characterize the alignment transforms under consideration. As the spatial and temporal factors are considered independently in this work, we are in a position to discuss them separately.

3.1   The Spatial Alignment Submanifold

The previous examples imply that the spatial alignment manifold $\mathfrak{M}_S$ is usually identical to a Riemannian manifold of the transformation/constraint matrices. Affine group $\mathbb{A}(2)$ and special linear group $\mathbb{SL}(3)$ both belong to the *matrix Lie group*, which possesses several intrinsic geometric properties. We list a few used in this work: the geodesic (intrinsic) distance between two elements $\mathbf{V}_1, \mathbf{V}_2$ on the matrix Lie group is $d(\mathbf{V}_1, \mathbf{V}_2) = \|\log(\mathbf{V}_1^{-1}\mathbf{V}_2)\|$. The exponential map $\mathscr{E}_{v_m} : \mathscr{T}_{v_m} \to \mathbb{G}$, which maps v' in the tangent space $\mathscr{T}_{\mathbf{V}_m}$ at $\mathbf{V}_m$ onto the group $\mathbb{G}$, is given by $\mathscr{E}_{\mathbf{V}_m}(\mathbf{V}') = \mathbf{V}_m \exp(\mathbf{V}_m^{-1}\mathbf{V}')$. The logarithmic map $\mathscr{L}_{\mathbf{V}_m} : \mathbb{G} \to \mathscr{T}_{\mathbf{V}_m}$, meanwhile, is $\mathscr{L}_{\mathbf{V}_m}(\mathbf{V}) = \mathbf{V}_m \log(\mathbf{V}_m^{-1}\mathbf{V})$. The matrix exponential and logarithmic operation used here are defined as $\exp(\mathbf{X}) = \sum_{i=0}^{\infty} \frac{1}{i!}\mathbf{X}^i$ and $\log(\mathbf{X}) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i}(\mathbf{X} - \mathbf{I})^i$.

The space of fundamental matrices - $\mathbf{F}$'s, as in Example 3, is the space of those matrices with rank 2. To get a parameterization for this manifold, we employ the singular value decomposition $\mathbf{F} = \mathbf{U}_1 \mathbf{\Sigma} \mathbf{U}_2^T$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are both 3×2 orthogonal matrices and $\mathbf{\Sigma}$ is 2×2 diagonal positive. It is known that the spaces of all 3×2 orthogonal matrices is Stiefel manifold $\mathfrak{V}_{2,3}$ [14] and thus the spatial alignment manifold $\mathfrak{M}_S = \mathfrak{V}_{2,3} \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathfrak{V}_{2,3}$. For two elements $\mathbf{V}_1, \mathbf{V}_2$ on $\mathfrak{V}_{2,3}$, an intrinsic distance is $d(\mathbf{V}_1, \mathbf{V}_2) = \sqrt{2 - tr(\mathbf{V}_1^T\mathbf{V}_2)}$. The tangent vectors at $\mathbf{V}_m$, denoted as $\mathbf{V}'$'s, can be represented as $\mathbf{V}' = \mathbf{V}_m\mathbf{A} + (\mathbf{I} - \mathbf{V}_m\mathbf{V}_m^T)\mathbf{B}$, where $\mathbf{A}$ is skew-symmetric and $\mathbf{B}$ is arbitrary. The exponential map from $\mathbf{V}'$ to $\mathbf{V}$, meanwhile, can be obtained as

$$\mathbf{V} = [\mathbf{V}_m, \mathbf{Q}] \exp\left(\begin{bmatrix} \mathbf{V}_m^T\mathbf{V}' & -\mathbf{R}^T \\ \mathbf{R} & \mathbf{0} \end{bmatrix}\right)\begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \tag{3}$$

where $\mathbf{Q}$ and $\mathbf{R}$ are the QR-decomposition of $(\mathbf{I} - \mathbf{V}_m\mathbf{V}_m^T)\mathbf{V}'$.

3.2 The Temporal Alignment Submanifold

As pointed out earlier, in this work we not only account for the temporal mis-alignment due to synchronization problem and differences in frame rates of the cameras, but also exploit the rate variations within observed dynamic instances of the same category. Rate variation within a fixed time span, *i.e.*,[0,1], with global frame rate (scaling) and shift eliminated, is well modeled as a diffeomorphism γ from [0,1] to [0,1] with $\gamma(0) = 0$ and $\gamma(1) = 1$ [11]. Then, any time warping or misalignment $W(t)$ under consideration can be written as $W(t) = k_2\gamma(\frac{t-l_1}{k_1}) + l_2$, where k_1, k_2 are the positive global scaling factors and l_1, l_2 are the shift factors, defined for $l_1 \leq t \leq k_1 + l_1$. Obviously, when we take $\gamma(t) = t$, $W(t)$ reduces to the temporal affine transformation. Denoting the space of all possible γ's as $\mathfrak{d}$, we can now formally define the temporal alignment submanifold as $\mathfrak{M}_T = \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R} \times \mathfrak{d}$, where $\mathbb{R}^+ \times \mathbb{R}^+$ accounts for k_1, k_2 and $\mathbb{R} \times \mathbb{R}$ accounts for l_1, l_2.

If we let $\psi = \sqrt{\dot{\gamma}}$ and the space of all ψ's to be $\ominus$, then under Fisher-Rao metric (See [15, 16]), the intrinsic distance between ψ_1 and ψ_2 are $d(\psi_1, \psi_2) = \cos^{-1}(< \psi_1, \psi_2 >)$ where $< \psi_1, \psi_2 >= \int_0^1 \psi_1(t)\psi_2(t)dt$. The exponential map $\mathscr{E}_{\psi_m} : \mathscr{T}_{\psi_m} \to \ominus$ for $\psi' \in \mathscr{T}_{\psi_m}$ is defined as $\mathscr{E}_{\psi_m}(\psi') = \cos(< \psi', \psi' >^{\frac{1}{2}})\psi_m + \frac{\sin(<\psi',\psi'>^{\frac{1}{2}})}{<\psi',\psi'>^{\frac{1}{2}}}\psi'$. The logarithmic map $\mathscr{L}_{\psi_m} : \ominus \to \mathscr{T}_{\psi_m}$, which is actually the inverse map of exponential map, is then given by $\mathscr{L}_{\psi_m}(\psi) = \frac{\arccos(<\psi,\psi_m>)}{<\psi^*,\psi^*>^{\frac{1}{2}}}\psi^*$, where $\psi^* = \psi_m - < \psi, \psi_m > \psi$. Since we have used ψ instead of γ, the temporal alignment submanifold can also be equivalently represented as $\mathfrak{M}_T = \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R} \times \ominus$.

4 Sequential Importance Sampling on the Manifold for Optimal Alignment

It is now clear that the alignment problem (1) becomes an optimization problem on the alignment manifold $\mathfrak{M}$. This problem differs from previous works where exhaustive or greedy strategies are employed pertaining to a specific spatio-temporal parameter space, which is usually treated as Euclidean. Meanwhile, the gradient or Newton methods as used previously will tend to fall into local optimum as J defined on $\mathfrak{M}$ is normally non-convex and multi-modal. In sum, it is desirable to find an algorithm that accounts for the non-linear manifold of the arguments, converges to the global optimum, and has reasonable computational complexity.

Let us consider the following time-varying state-space model:

$$\begin{bmatrix} \mathbf{p}_{S,h} \\ \mathbf{p}_{T,h} \end{bmatrix} = \begin{bmatrix} \mathscr{E}_{\mathbf{p}_{S,h-1}}(\mathbf{u}_{S,h}) \\ \mathscr{E}_{\mathbf{p}_{T,h-1}}(\mathbf{u}_{T,h}) \end{bmatrix} \tag{4}$$

$$\mathbf{y}_h = J(S^1, S^2, \mathbf{p}_h) - v_h. \tag{5}$$

where $\mathbf{p}_h = [\mathbf{p}_{S,h}^T, \mathbf{p}_{T,h}^T]^T$ is the parameter state at step h. We assume that $\mathbf{p}^*$, the optimal alignment, is not directly observable, while at step t we observe $\mathbf{y}_t$. Moreover, we let $\mathbf{u}_{S,h} \sim \mathcal{N}(\mathbf{0}, (\sigma_S/h)^2\mathbf{I})$, $\mathbf{u}_{T,h} \sim \mathcal{N}(\mathbf{0}, (\sigma_T/h)^2\mathbf{I})$, where σ_S^2 and σ_T^2 are both small numbers. By construction (details below) we may let v_h to be an nonnegative random variable with an appropriate density function (*e.g.*, exponential $\mathcal{E}(\lambda)$ in this work). Equivalently, we may represent the state transition and observation model as $p(\mathbf{p}_{S,h}|\mathbf{p}_{S,h-1}) \sim \exp(-\frac{d^2(\mathbf{p}_{S,h}, \mathbf{p}_{S,h-1})}{2(\sigma_T/h)^2})$, $p(\mathbf{p}_{T,h}|\mathbf{p}_{T,h-1}) \sim \exp(-\frac{d^2(\mathbf{p}_{T,h}, \mathbf{p}_{T,h-1})}{2(\sigma_T/h)^2})$, where $p(\mathbf{p}_h|\mathbf{p}_{h-1}) \sim p(\mathbf{p}_{S,h}|\mathbf{p}_{S,h-1})p(\mathbf{p}_{T,h}|\mathbf{p}_{T,h-1})$, and $p(\mathbf{y}_h|\mathbf{p}_h) \sim \exp(\lambda(\mathbf{y}_h - J(S^1, S^2, \mathbf{p}_h)))$.

The motivation as to why we formulate a state space model is to be able to recursively compute the Maximum A Posterior (MAP) estimate of the parameter state $p(\mathbf{p}_h|\mathbf{y}_h, \mathbf{y}_{h-1}, \cdots, \mathbf{y}_0)$. From the recursion $p(\mathbf{p}_h|\mathbf{y}_h, \mathbf{y}_{h-1}, \cdots, \mathbf{y}_0) \propto p(\mathbf{y}_h|\mathbf{p}_h) \int p(\mathbf{p}_h|\mathbf{p}_{h-1})p(\mathbf{p}_{h-1}|\mathbf{y}_{h-1}, \mathbf{y}_{h-2}, \cdots, \mathbf{y}_0)d\mathbf{p}_{h-1}$, we know that the posterior probability of the alignment $p(\mathbf{p}_h|\mathbf{y}_h, \mathbf{y}_{h-1}, \cdots, \mathbf{y}_0)$ is equal to the posterior probability at the previous step $p(\mathbf{p}_{h-1}|\mathbf{y}_{h-1}, \mathbf{y}_{h-2}, \cdots, \mathbf{y}_0)$ smoothed by the state transition probability $p(\mathbf{p}_h|\mathbf{p}_{h-1})$ and weighted by the likelihood $p(\mathbf{y}_h|\mathbf{p}_h)$. Therefore, by constructing a decreasing sequence $\{\mathbf{y}_h\}_{h=0,1,\dots}$ and letting σ_S, σ_T be small, $p(\mathbf{p}_h|\mathbf{y}_h, \mathbf{y}_{h-1}, \cdots, \mathbf{y}_0)$ is expected to be continuously increasing and peaking at the the optimal alignment $\mathbf{p}^*$. In other words, the MAP estimate of the parameter state will give the optimal alignment.

The above Bayesian recursive estimation is realized in a Monte Carlo manner. In particular, the construction of appropriate observation sequence $\{\mathbf{y}_h\}_{h=0,1,\dots}$ come up naturally from the Monte Carlo samples. We propose the SIS algorithm on the alignment manifold as follows. Note that the proposed algorithm handles states evolving on the Riemannian manifold rather than the conventional Euclidean space, thus is different from most existing particle filters and their variations. Bayesian recursive filtering using particles has been proposed for specific manifolds in the context of tracking [17–20], while the following approach is generally applicable for various alignment manifolds. Furthermore, we formulate the static optimization problem into a dynamic state space model, which provides insight on applications of SIS to new problems beyond tracking.

Algorithm *SIS on the alignment manifold.*

1)Initialization. Specify an initial distribution p_0 defined on $\mathfrak{M}$ and draw i.i.d. samples $\{\mathbf{p}_0^k\}_{k=1}^K$ from p_0. Let $h = 1$.

2)Importance Sampling. Sample $\hat{\mathbf{p}}_h^k$ from $p(\mathbf{p}_h^k|\mathbf{p}_{h-1}^k)$. For this purpose, generate $\mathbf{u}_{S,h}^k$ from $\mathcal{N}(\mathbf{0}, (\sigma_S/h)^2\mathbf{I})$ and $\mathbf{u}_{T,h}^k$ from $\mathcal{N}(\mathbf{0}, (\sigma_T/h)^2\mathbf{I})$. Then apply exponential maps $\hat{\mathbf{p}}_{S,h}^k = \mathscr{E}_{\mathbf{p}_{S,h-1}^k}(\mathbf{u}_{S,h}^k)$ and $\hat{\mathbf{p}}_{T,h}^k = \mathscr{E}_{\mathbf{p}_{T,h-1}^k}(\mathbf{u}_{T,h}^k)$.

3)Constructing observation. Let

$$\mathbf{y}_h = \min_k J(S^1, S^2, \hat{\mathbf{p}}_h^k). \tag{6}$$

If $\mathbf{y}_h > \mathbf{y}_{h-1}$, $\mathbf{y}_h \longleftarrow \mathbf{y}_{h-1}$.

4)Weighting. Approximate the new posterior probability by

$$q_h(\mathbf{p}_h) = \sum_{k=1}^{K} w_h^k \delta(\mathbf{p}_h - \hat{\mathbf{p}}_h^k), \qquad (7)$$

where δ is the Kronecker delta, $w_h^k \propto p(\mathbf{y}_h|\hat{\mathbf{p}}_h^k)$ and $\sum_{k=1}^{K} w_h^k = 1$.

5)Importance resampling. Draw i.i.d. samples $\{\mathbf{p}_h^k\}_{k=1}^{K}$ from $q_h(\mathbf{p}_h)$.

6)Stop if a stopping criteria is satisfied; Otherwise, $h \longleftarrow h+1$ and go to 2).

Step 3) follows from the observation equation in the proposed state-space model, and this construction of observation $\mathbf{y}_h$ plays an important role in the above algorithm. By letting $\mathbf{y}_h$ to be the minimum value of the alignment cost function, Monte Carlo samples that lead to a lower cost will receive higher importance weights when applying the weighting step. Consequently, the Monte Carlo samples (particles) will tend to concentrate around the minima of the alignment cost function, including the global minimum. With a proper initialization of samples over $\mathfrak{M}$, the optimal $\mathbf{p}^*$ will be located more and more accurately during the coarse-to-fine particle propagation. The operation $\mathbf{y}_h \longleftarrow \mathbf{y}_{h-1}$ when $\mathbf{y}_h > \mathbf{y}_{h-1}$ guarantees non-increasing $\mathbf{y}_h$.

The initialization of the particles is case dependant. As an example for the spatial alignment submanifold of $\mathbb{A}(2)$, we may generate independent, uniformly distributed samples over the corresponding Lie algebra $\partial(2)$ and exponentially map them onto $\mathbb{A}(2)$. For the temporal alignment submanifold, we may also generate uniform distributed samples over the tangent space at $\gamma(t) = t$ together with uniform samples from $\mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}$. The stopping criteria, meanwhile, can be flexible as well. $0 < \mathbf{y}_{h-1} - \mathbf{y}_h < \epsilon$ is a useful one. The final MAP estimate of $\mathbf{p}^*$, can be simply taken as $\hat{\mathbf{p}}^* = \arg\min_k J(S^1, S^2, \hat{\mathbf{p}}_h^k)$ after the algorithm stops at step h.

5   Empirical Evaluation

We have applied the algorithm described above to three different datasets for the same purpose of spatial-temporal alignment, while these datasets represent different spatio-temporal signals originated from videos. Specifically, we looked into the alignment of point trajectories, deforming shape sequences, as well as videos themselves. The alignment objectives and alignment manifolds corresponding to each datasets vary, while the SIS procedure is the same for all. In each experiment, we select appropriate state-of-art methods or design baseline(s) for comparison, while the purpose of these comparisons is simply to show how the inclusion of temporal warping submanifold, formulation of the aligning procedure as a recursive estimation of the state-space model, and the Monte Carlo approach help advance the state-of-art performance on practical data.

5.1   Evaluation with Point Trajectories

We first evaluate our method with point trajectories, which are essentially the input to feature-based methods. In this paper we make use of the GaTech Football Multi-Trajectory Dataset [21, 22]. This dataset contains 55 sets of trajectories and in each set there are eleven trajectories corresponding to the movements of the eleven offensive players in a play. The sets are organized into categories, each of which contains all realizations of the same play strategy (specified by the playbook). In other words, trajectory sets in the same category are samples of the same 'activity', thus resembling each other (on the ground plane) though intra-category variations exist. However, they are observed in different viewpoint and executed at different and varying rates. In each set the roles of players are annotated and thus the trajectory correspondence between two sets is available to us.

We model the spatial misalignments to be a planar homography and thus the spatial alignment submanifold becomes $\mathbb{SL}(3)$. The misalignment cost J is simply taken as the average distances between point pairs from all trajectory pairs across the whole time span. We perform two types of experiments, in the first of which we select a set of trajectories and transform it with a typical view change (homography) and a specific time warping to get the other, and then we align the two. We do so on all 55 sets. In the second type, we randomly select a total of 40 pairs of sets, each pair being the samples of the same play type (activity), and then we align these pairs. For comparative purposes, we implemented two state-of-art methods [7, 3] that address similar task as ours. The approach in [7] assumes affine temporal misalignment only, and the strategy in [3] uses Dynamic Time Warping (DTW) to determine the non-linear temporal misalignment. The preprocessing modules of tracking and correspondence in the two methods are unnecessary as the dataset has provided trajectory and correspondence information, and thus a common basis is shared among all implementations for comparison. Note that [7] mainly focus on temporal alignment, and to add spatial alignment into it we simply estimate a planar homography with the points from the temporally aligned trajectories. Meanwhile, when using [3] we take alternations between DTW and gradient-descent-based homography estimation (on all corresponding points collected from all temporally aligned frames) to get the final alignment parameters. (Note that though DTW is globally optimal in 1-D temporal dimension, when placed into alternations between spatial and temporal submanifolds the combined search may not necessarily be so, and thus the alternating process is a greedy search.) For our method, we get the initial particles by generating random samples in the tangent space at the homography estimated from the first pair frames and in the tangent space at $\gamma(t) = t$.

Samples of the results are shown in Figure 1, where in the first two columns are the two trajectory sets to be aligned toward each other, and the following two columns show alignment results using the state-of-art methods and our method. Each of the three rows, meanwhile, represents a typical experimental setting: in row (a) the target is a generated misaligned version of the reference, in row (b)

the reference and target are a real pair with similar realization but undergoing significant misalignment, and in row (c) the two are a real pair with significant variation from each other.
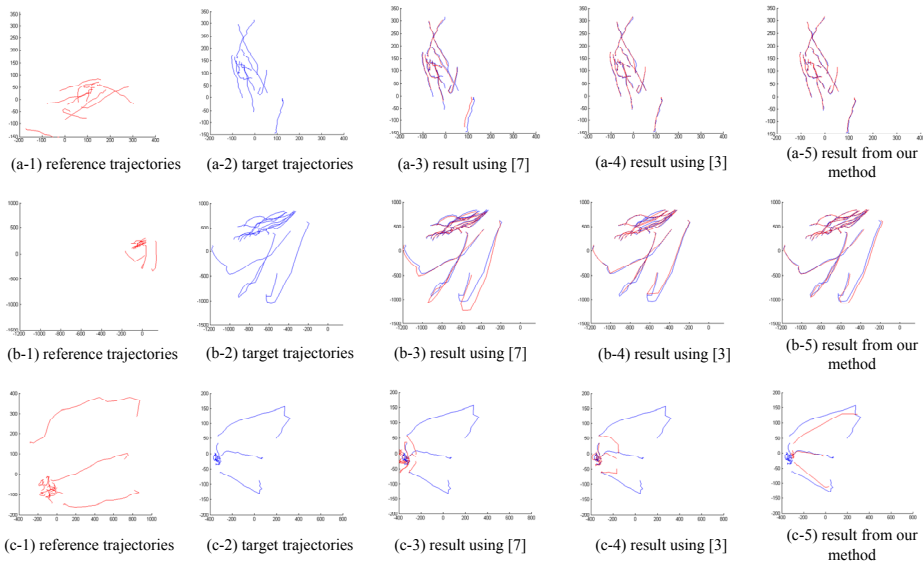

Fig. 1. Samples of the alignment results on point trajectories.

To quantitatively understand the performance of the alignment methods, we recorded the average distance of point pairs from aligned trajectory pairs, and show the results in Table 1. Note that the statistics is from the 40 real pairs rather than the generated ones.

Table 1. Average residual misalignments between the aligned trajectory pairs.

	mean	standard deviation
Using [7]	15.9	8.6
Using [3]	13.1	6.8
Our method	10.0	3.6

5.2   Evaluation with Deforming Shape Sequences

Sequences of deforming shapes are typical mid-level features extracted from original videos containing the deforming objects of interest. In this experiment we

use silhouette sequences from the USF Gait Database [23] to demonstrate the performance of out method. We randomly select 20 sequence pairs, each with the same shoe types, carrying conditions, surface types, and walking directions, but observed at two different times. For efficiency, in each sequence we only consider the segment of frames of the first two walking circles. The spatial misalignment within each pair is modeled as affine and is actually less significant compared to the GaTech Football Multi-Trajectory Dataset, and the main focus is on the effect of taking non-linear rate warping into account in addition to linear scaling and shift. For comparison, we implemented the gradient descent algorithm presented in [5] and designed one more baseline. The designed baseline alternates between DTW (on all frames from spatial alignment) and gradient-descent-based affine estimation (on all frame pairs temporally aligned), and thus is a greedy search. The cost function is simply taken as the sum of pixel-wise absolute differences. For [5], the initial spatial parameter is estimated as a translation between the leading frames and the initial temporal parameter is taken as $\gamma(t) = t$. For our method, Monte Carlo samples are generated from Gaussians in the tangent spaces at the initial parameters.

We show two sample results in Figure 2, where each of the five rows for each sequence pair is explained in the caption of the figure. The average residual misalignment errors (in pixels) for all 20 pairs are shown in Table 2. Note that all three methods perform well due to mild spatial misalignment and near-affine temporal misalignment, while our method achieves improvement over [5] by allowing non-linear warping effect, and the improvement over alternating DTW and affine estimation should be credited to better global convergency.

Table 2. Average residual misalignments between the pairs of shape sequences.

	mean	standard deviation
Using [5]	23.5	6.4
Alternating DTW and affine estimation	26.7	6.8
Our method	21.3	7.1

5.3   Evaluation with Human Action Videos

In the third set of experiments, we work with human action videos directly. We use the KTH database [24], in which the semantically meaningful signal is human motion. We randomly select 30 pairs of sequences, each pair performing the same action, but moderate variations in clothing, background, or view angle exist within the pair. For efficiency, again for each sequence we only keep a segment of frames including human motion but discard pure background frames. The spatial misalignment within each pair is affine [10], and the misalignment cost is the spatio-temporal correlation used by [10] but on optical flow extracted from consecutive frames. For comparison, we implemented [10] and the method that alternates between DTW and affine estimation as in previous section. The
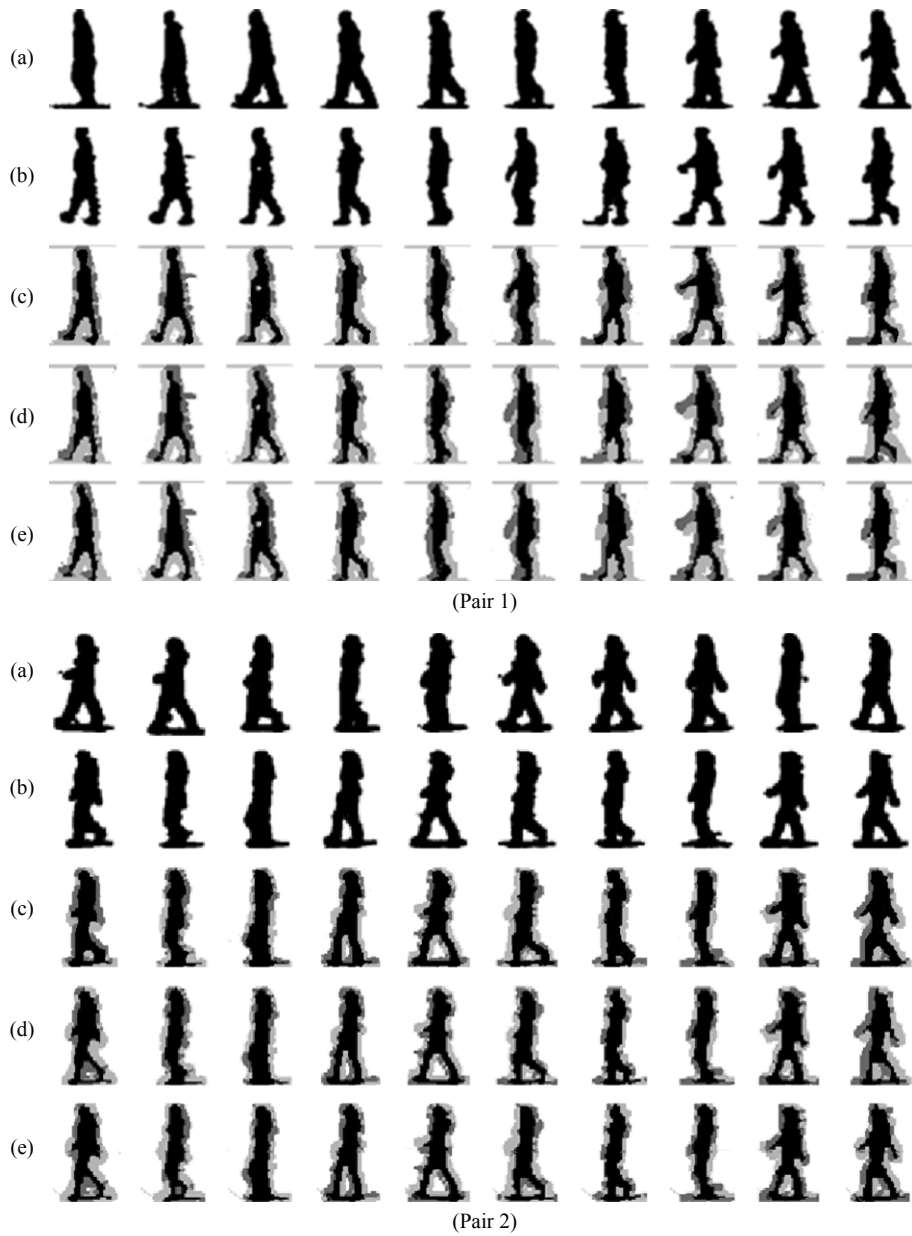
(a)

(b)

(c)

(d)

(e)

(Pair 1)

(a)

(b)

(c)

(d)

(e)

(Pair 2)

Fig. 2. Samples of the alignment results on deforming shape sequences from USF Gait Database. For each pair, (a) is the reference sequence and (b) is the target. (c), (d), and (e) give the alignment results (transformed sequence overlaid onto target) using our method, the method in [5], and the method that alternates between DTW and spatial alignment. The black, dark shaded, light shaded, and white areas denote true positive, false negative, false positive and true negative respectively. In other words, the black and dark shaded areas constitute the silhouette of the target, while the black and lighted shaded areas constitute the transformed silhouette. Therefore, a larger black area implies a better alignment.

initial spatial parameter, when necessary, is estimated as the translation between the leading frames and the initial temporal parameter is taken as $\gamma(t) = t$. Meanwhile, Monte Carlo samples are generated from Gaussians in the tangent spaces at the initial parameters too.

We show three sample results in Figure 3, where each of the five rows for each sequence pair has the same interpretation as in the previous section. Substantial execution rate variations exist within every pair, and changes in clothing, background, or view angle also exist. There is not a numerical criterion to evaluate the performance on aligning real videos, and by qualitative observation the proposed method performs comparatively well as the baselines, and is visually more close to the target when undergoing a larger view change (pair 3).

As is true for many efforts involving particle filters, the proposed method is computationally more demanding than greedy search, but much less expensive than exhaustive approaches. This trade-off, however, leads to improved performance as demonstrated in the previous subsections. The time complexity depends on the number of particles used. The convergence, on the other hand, turns out to be fast. In this section all results are obtained with 1000 particles and less than twenty iterations. Another issue is that $\mathfrak{b}$ is by definition infinitely dimensional, while in all experiments we approximated the γ's with non-decreasing sequences valued from 0 to 1 of length 20.

6   Discussion

This work assumes that the parametric manifolds are known a priori; alignment problems without knowing the specific form of the manifolds deserve exploration as well. It is also desirable to remove the assumption regarding relative stationarity between cameras. Though we pursue global optimum in the algorithm and empirically observe improved solution, we have not theoretically proved any properties regarding asymptotic convergence. A theoretical study on geometrical SIS method will be important. We will also look into other efficient search schemes like stochastic gradient descent. By generalizing the considered manifolds and cost functions, we will extend the proposed strategy of stochastic optimization on geometric spaces for other problems (*e.g.* face alignment on Grassmann manifold [25]). We hope this can bring new insights and improved performance to a larger number of vision applications.

References

1. Lee, L., Romano, R., Stein, G.: Monitoring activities from multiple video streams: Establishing a common coordinate frame. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 758–767
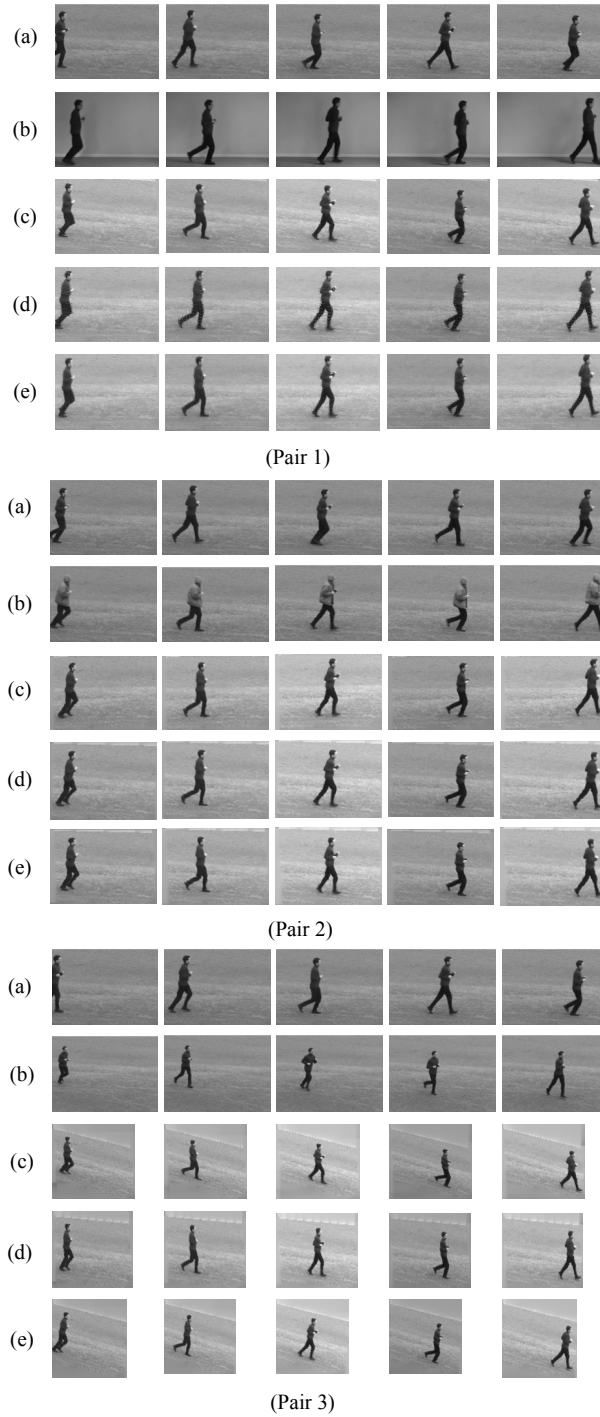2. Wolf, L., Zomet, A.: Sequence to sequence self calibration. In: ECCV. (2002)

Fig. 3. Samples of the alignment results on KTH dataset. For each pair, (a) is the reference sequence and (b) is the target. (c), (d), and (e) give the alignment results (transformed sequence overlaid onto target) using our method, the method in [10], and the method of alternation between DTW and spatial alignment.

3. Rao, C., Gritaiand, A., Shah, M., Syeda-Mahmood, T.: View-invariant alignment and matching of video sequences. In: ICCV. (2003)
4. Laptev, I., Belongie, S., Perez, P., Wills, J.: Periodic motion detection and segmentation via approximate sequence alignment. In: ICCV. (2005)
5. Caspi, Y., Simakov, D., Irani, M.: Feature-based sequence-to-sequence matching. International Journal of Computer Vision **68** (2006) 53–64
6. Wolf, L., Zomet, A.: Wide baseline matching between unsynchronized video sequences. International Journal of Computer Vision **68** (2006) 43–52
7. Padua, F., Carceroni, R., Santos, G., Kutulakos, K.: Linear sequence-to-sequence alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010) 304–320
8. Caspi, Y., Irani, M.: A step towards sequence-to-sequence alignment. In: CVPR. (2000)
9. Caspi, Y., Irani, M.: Alignment of non-overlapping sequences. In: ICCV. (2001)
10. Ukrainitz, Y., Irani, M.: Aligning sequences and actions by maximizing space-time correlations. In: ECCV. (2006)
11. Veeraraghavan, A., Srivastava, A., Roy-Chowdhury, A., Chellappa, R.: Rate-invariant recognition of humans and their activities. IEEE Transactions on Image Processing **18** (2009) 1326–1339
12. Zhou, F., de la Torre, F.: Canonical time warping for alignment of human behavior. In: NIPS. (2009)
13. Gordon, N., Salmond, D., Smith, A.: Novel approach to nonlinear/non-gaussian bayesian state estimation. IEE Proceedings F on RAdar and Signal Processing **140** (1993) 107–113
14. Arias, T., Edelman, A., Smith, S.: The geometry of algorithms with orthogonality constraints. SIAM Journal of Matrix Analysis and Applications **20** (1998) 303–353
15. Maybank, S.: The Fisher-Rao metric for projective transformations of the line. International Journal of Computer Vision **63** (2005) 191–206
16. Srivastava, A., Jermyn, I., Joshi, S.: Riemannian analysis of probability density functions with applications in vision. In: CVPR. (2007)
17. Srivastava, A., Klassen, E.: Bayesian and geometric subspace tracking. Advances in Applied Probability **36** (2004) 43–56
18. Wu, Y., Wu, B., Liu, J., Lu, H.: Probabilistic tracking on riemannian manifolds. In: ICPR. (2008)
19. Kwon, J., Lee, K.M., Park, F.C.: Visual tracking via geometric particle filtering on the affine group with optimal importance functions. In: CVPR. (2009)
20. Porikli, F., Pan, P.: Refressed importance sampling on manifolds for efficient object tracking. In: AVSS. (2009)
21. Li, R., Chellappa, R., Zhou, S.K.: Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition. In: CVPR. (2009)
22. Li, R., Chellappa, R.: Group motion segmentation using a spatio-temporal driving force model. In: CVPR. (2010)
23. Sarkar, S., Phillips, P.J., Liu, Z., Robledo, I., Grother, P., Bowyer, K.W.: The human id gait challenge problem: Data sets, performance, and analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 162–177
24. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR. (2004)
25. Lui, Y.M., Beveridge, J.R.: Grassmann registration manifolds for face recognition. In: ECCV. (2008)