Initial-State Invariant Binet-Cauchy Kernels for the Comparison of Linear Dynamical Systems

Rizwan Chaudhry[1] René Vidal[2]

***Abstract*— Linear Dynamical Systems (LDSs) have been extensively used for modeling and recognition of dynamic visual phenomena such as human activities, dynamic textures, facial deformations and lip articulations. In these applications, a huge number of LDSs identified from high-dimensional time-series need to be compared. Over the past decade, three computationally efficient distances have emerged: the Martin distance [1], distances obtained from the subspace angles between observability subspaces [2], and distances obtained from the family of Binet-Cauchy kernels [3]. The main contribution of this work is to show that the first two distances are particular cases of the latter family obtained by making the Binet-Cauchy kernels invariant to the initial states of the LDSs. We also extend Binet-Cauchy kernels to take into account the mean of the dynamical process. We evaluate the performance of our metrics on several human activity recognition datasets and show similar or better results.**

I. Introduction

Linear Dynamical Systems (LDSs) have been extensively used for modeling and recognition of dynamic visual phenomena. For instance, [4] uses LDSs to model surgical gestures in video data from the DaVinci robot; [5], [6] use LDSs to model the appearance of a deforming heart in a magnetic resonance image sequence; [7], [8], [9], [10], [11], [12], [13] use LDSs to model the appearance of *dynamic textures*, such as water or fire, in a video sequence; [14], [15], [16], [17], [18], [19] use LDSs to model human gaits, such as walking or running, in motion capture and video data; [20] uses LDSs to model the appearance of moving faces; and [21] uses LDSs to model audio-visual lip articulations.

In these applications, the recognition pipeline consists of the following steps: 1) extract a time-series of appropriate features, 2) model the time-series using LDSs, 3) compute a metric between dynamical systems and 4) use algorithms such as Nearest-Neighbors or SVMs for classification. Arguably, the most important step is 3), which requires computationally efficient distances for comparing a huge number of LDSs identified from high-dimensional time-series data.

Related work. Existing methods for comparing LDSs can be broadly divided into three (sometimes overlapping) main categories: (1) Riemannian distances on spaces of LDSs, (2) metrics on their power spectra, and (3) metrics induced from a metric in a suitable ambient space.

*This work was supported in part by NSF grants 0941362, 0941463, 0931805, and 1335035

[1]R. Chaudhry is with Microsoft Corporation, Mountain View, CA 94043, USA, Rizwan.Chaudhry@microsoft.com

[2]R. Vidal is with the Center for Imaging Science, Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA, rvidal@jhu.edu

Methods in the first category were studied in the 70's and 80's for applications in system identification. [22], [23], [24], [25] deal explicitly with defining distances and study the geometrization of the smooth manifold of LDSs of fixed McMillan degree and size. Interestingly, for most other spaces of LDSs (e.g., LDSs of fixed size and McMillan degree not larger than a fixed number or arbitrary McMillan degree), a smooth finite dimensional Riemannian structure does *not* exist. However, this quotient geometry approach is limited to deterministic systems and the huge cost needed to actually compute a distance is not addressed [22], [23].

Methods in the second category compare two LDSs with output dimension p by comparing the power spectra. For example, one can use a matrix-norm-based distance on the *infinite dimensional* space of $p \times p$ spectral density matrices, $\mathcal{P}_p$. This distance can also be derived from the so-called Wasserstein distance between processes [26]. Other approaches consider a smaller subspace $\mathcal{P}_p^+$ of *full-rank* $p \times p$ spectral density matrices. Due to the strict positive-definiteness of spectral density matrices, $\mathcal{P}_p^+$ naturally has the structure of an infinite dimensional open cone. Amari in [27], [28] gives an infinite dimensional Riemannian and a more general information geometry based framework to geometrize $\mathcal{P}_p^+$ mainly for the case of $p = 1$ (see also [29]). Amari's framework can be extended to $p > 1$ and in that direction recently an infinite-dimensional Riemannian framework [30] has been suggested. However, key disadvantages of these methods for large p are that they are computationally expensive and the assumption of full-rankness is too restrictive and not realistic.

Methods in the third category include metrics based on subspace angles [2] such as the Martin distance [1], algebraic metrics such as the Binet-Cauchy kernels [3], the alignment distance [31], and probabilistic metrics such as the KL-divergence [32]. Of these, the Martin distance has been the most extensively used as it is invariant to the noise statistics as well as initial state of the dynamical system. For human activity recognition, the initial state is usually not relevant. For example, we do not want to discriminate between a walking action which is observed starting mid-cycle with both feet crossing each other and a walking action where the two feet are at opposite ends. The original Binet-Cauchy kernel in [3], however, is not invariant to the initial state of the dynamical system and therefore does not perform very well in these tasks. While it is possible to make the Binet-Cauchy kernel initial-state invariant by taking the expectation over the initial states, as proposed in [3], this approach has not been explored due to the difficulty in defining an

appropriate distribution for the initial states.

Paper contributions. We propose two approaches to making the Binet-Cauchy kernels initial-state invariant. The first approach is based on computing the determinant of a matrix relating the two dynamical models. We prove that this kernel is invariant to transformations of the dynamical models. We also show that the Martin kernel is a particular case of the proposed determinant kernel. The second approach is based on maximizing the Binet-Cauchy kernel with respect to the initial states. We show that all metrics based on subspace angles are a particular case of this approach. We also extend Binet-Cauchy kernels to take into account the mean of the dynamical process. We extensively test our proposed metrics against the Martin distance and the original Binet-Cauchy kernel for the task of human activity recognition and show that we get superior recognition performance.

Paper outline. The rest of the paper is organized as follows. §II gives the relevant technical background for dynamical systems-based modeling as well as briefly summarizes the original Binet-Cauchy kernel. In §III, we propose two initial-state invariant Binet-Cauchy kernels and provide theoretical results relating them with existing metrics for dynamical systems. In §IV, we extend our kernels to take into account the mean of the dynamical process. In §V, we provide the results of several human activity recognition experiments to display the efficacy of our proposed metric. Finally, in §VI, we provide conclusions and directions for future research.

II. Background

This section provides a brief overview of dynamical systems and metrics on the space of dynamical systems. We limit our review to distances that are computationally efficient for high-dimensional systems.

A. Dynamical systems

Given a time-series, $\{\mathbf{y}_t \in \mathbb{R}^p\}_{t=1}^T = [\mathbf{y}_1, \dots, \mathbf{y}_T]$, a Linear Dynamical System (LDS) models its temporal evolution using the following Gauss-Markov process:

$$\begin{aligned} \mathbf{x}_{t+1} &= A\mathbf{x}_t + B\mathbf{v}_{t+1} \\ \mathbf{y}_t &= \mu + C\mathbf{x}_t + \mathbf{w}_t. \end{aligned} \tag{1}$$

Here $\mathbf{x}_t \in \mathbb{R}^n$ represents the internal (hidden) state of the LDS at each time instant t, n represents the *order* of the LDS, $A \in \mathbb{R}^{n\times n}$ represents the *dynamics* matrix that linearly relates the states at time instants t and $t+1$, $C \in \mathbb{R}^{p\times n}$ represents the *observation* matrix that linearly transforms the internal state to the output $\mathbf{y}_t$, $\mu \in \mathbb{R}^p$ represents the mean of the output time-series. $\mathbf{v}_t \in \mathbb{R}^n$ and $\mathbf{w}_t \in \mathbb{R}^p$ correspond to the input and output noise processes usually assumed to be Gaussian with zero-mean. Specifically, $B\mathbf{v}_t \sim \mathcal{N}(0, Q)$, where $Q = BB^\top$, and $\mathbf{w}_t \sim \mathcal{N}(0, R)$, where $R = \sigma^2 I$. Given the time-series $\{\mathbf{y}_t\}_{t=1}^T$, the task of computing the system parameters, $(\mathbf{x}_0, \mu, A, C, B, R)$, is referred to as system identification and several optimal [33], [34] and sub-optimal but very efficient [7] methods have been proposed in literature. Chan et al. [35] proposed an extension of LDS to kernel Non-Linear Dynamical Systems (NLDS) by implicitly embedding a non-Euclidean time-series into a reproducing kernel Hilbert space (RKHS) using an appropriate kernel on the original non-Euclidean space.

B. Metrics for dynamical systems

Given a pair of LDS, $\mathcal{M}_i = (\mathbf{x}_{0;i}, \mu_i, A_i, C_i, B_i, R_i)$ for $i = 1, 2$, existing recognition algorithms define a metric between them, $d(\mathcal{M}_1, \mathcal{M}_2)$, for the purpose of comparison. As we have mentioned in the introduction, several metrics have been proposed in literature.

Martin distance [1], [2]. The Martin distance compares only the parameters A and C of the dynamical models. Let $\mathcal{M}_i = (A_i, C_i)$ for $i = 1, 2$. Assuming that the systems are stable, i.e., $\|A_i\|_2 < 1$, the Martin distance is defined as,

$$d_M(\mathcal{M}_1, \mathcal{M}_2)^2 = -\ln \prod_{i=1}^{n} \cos^2 \theta_i. \tag{2}$$

Here, θ_i is the i-th subspace angle between the range spaces of the infinite observability matrices O_1 and O_2 defined as

$$O_i = [C_i^\top, (C_i A_i)^\top, (C_i A_i^2)^\top, \dots] \quad \text{for} \quad i = 1, 2. \tag{3}$$

To compute the subspace angles we first solve the Sylvester equations $P_{ij} = A_i^\top P_{ij} A_j + C_i^\top C_j$ for $i, j = 1, 2$. We then compute the eigenvalues, $\{\lambda_i\}_{i=1}^{2n}$ of $\begin{bmatrix} 0 & P_{11}^{-1}P_{12} \\ P_{22}^{-1}P_{21} & 0 \end{bmatrix}$. The subspace angles, $\{\theta_i\}_{i=1}^n$ can then be computed as $\theta_i = \cos^{-1}(\lambda_i)$.

Binet-Cauchy kernels. Vishwanathan et al. [3] introduced an algebraic approach to comparing two LDS, leading to a complete family of kernels called the *Binet Cauchy* kernels. One of the proposed kernels, the Binet-Cauchy *trace* kernel between two LDS with uncorrelated noise processes, depends only on the parameters (A, C) of the LDS and the initial condition $\mathbf{x}_0$. Let $\mathcal{M}_i = (\mathbf{x}_{0;i}, A_i, C_i)$ for $i = 1, 2$. The trace kernel is defined as,

$$k_{tr}(\mathcal{M}_1, \mathcal{M}_2) = \mathbf{x}_{0;1}^\top P_{12} \mathbf{x}_{0;2}, \tag{4}$$

where P_{12} is the solution to the Sylvester equation,

$$P_{12} = \lambda A_1^\top P_{12} A_2 + C_1^\top C_2, \tag{5}$$

which is given by

$$P_{12} = \sum_{t=0}^{\infty} \lambda^t (A_1^t)^\top C_1^\top C_2 A_2^t. \tag{6}$$

This matrix exists and is unique if $\lambda \|A_1\|_2 \|A_2\|_2 < 1$.

An important property of the trace kernel is that it can be used to compare unstable systems (i.e., $\|A_i\|_2 > 1$) by choosing λ small enough. Moreover, the trace kernel is invariant with respect to a change of basis of the state space $x'_{t;i} = T_i x_{t;i}$. That is, $x'^\top_{t;i} P'_{ij} x'_{t;j} = x_{t;i} P_{ij} x_{t;j}$, where

$$(\mathbf{x}'_{0;i}, A'_i, C'_i) = (T_i \mathbf{x}'_{0;i}, T_i A_i T_i^{-1}, C_i T_i^{-1}). \tag{7}$$

The invariance property follows because

$$P'_{ij} = \lambda (T_i A_i T_i^{-1})^\top P'_{ij} (T_j A_j T_j^{-1}) + (C_i T_i^{-1})^\top (C_j T_j^{-1}),$$

and so $P'_{ij} = T_i^{-\top} P_{ij} T_j^{-1}$.

Extensions to NLDS. For the case of comparing non-Euclidean time-series, Chan et al. [35] and Chaudhry et al. [18] proposed the Martin distance and the Binet-Cauchy kernels for kernel NLDS respectively.

Invariance properties. As we can see, the Martin distance is invariant to the initial states of the dynamical system as well as the noise statistics. On the other hand, the Binet-Cauchy kernel is not invariant to these. Even in the case of uncorrelated noise as in Equation (4), the computation of the Binet-Cauchy kernel involves the initial states of the two LDS.

One way of making the Binet-Cauchy kernels invariant to initial conditions is to use take the expectation of the kernel with respect to the initial conditions of both systems, as proposed in [3]. More specifically, if $\Sigma_{\mathbf{x}_0} = \mathbb{E}(\mathbf{x}_{0;1}\mathbf{x}_{0;2}^\top)$ and $\mathcal{M}_i = (A_i, C_i)$ for $i = 1, 2$, the initial-state invariant Binet-Cauchy trace kernel with uncorrelated noise processes is defined as,

$$k_{tr}(\mathcal{M}_1, \mathcal{M}_2) = \text{trace}(\Sigma_{x_0} P_{12}). \tag{8}$$

However, $\Sigma_{\mathbf{x}_0}$ is not always available or deducible from the data. Hence there is a need to develop methods that do not require any statistics of the initial conditions.

III. Initial-state invariant Binet-Cauchy kernels

In this section, we propose two approaches to making the Binet-Cauchy kernel invariant with respect to the initial conditions. The first approach (described in §III-A) is based on computing a scalar function of P_{12}. In particular, we show that the *determinant kernel*, $\det(P_{12})$, is a positive-semidefinite kernel. We also show that a normalized version of it is invariant with respect to a change of basis. In addition, we show that the Martin kernel [1] is a particular case of the normalized determinant kernel. The second approach (described in §III-B) is based on maximizing the Binet-Cauchy kernel with respect to the initial conditions. We show that different maximizations lead to the different subspace angles [2] between LDSs. Hence, any metric based on subspace angles can be derived from the Binet-Cauchy kernel.

A. The determinant kernel

A simple approach to making the Binet-Cauchy kernel independent of the initial states is to consider any scalar function of P_{12}. For instance, we can consider the maximum singular value $k_{\max}(\mathcal{M}_1, \mathcal{M}_2) = \sigma_1(P_{12})$ or the trace $k_t(\mathcal{M}_1, \mathcal{M}_2) = \text{trace}(P_{12})$. However, it is not clear if these choices lead to a positive-definite kernel. Moreover, it is easy to see that such extensions are not invariant with respect to a change of basis.

In this section, we propose the following initial-state invariant extension of the Binet-Cauchy kernel:

$$k_d(\mathcal{M}_1, \mathcal{M}_2) = \det(P_{12})^2, \tag{9}$$

which we call the *Binet-Cauchy determinant kernel*[1]. The following theorem shows that this kernel is positive definite.

Theorem 1: k_d is a positive-definite kernel.

Proof: To show that k_d is a positive-definite kernel, we need to show that it can be written as $k_d(\mathcal{M}_1, \mathcal{M}_2) = \phi(\mathcal{M}_1)^\top \phi(\mathcal{M}_2)$ for some embedding ϕ. We construct such an embedding by making use of the Binet-Cauchy theorem for operators [3]. Specifically, [3] shows that for any operators F_1 and F_2 of compatible dimensions such that $F_1^\top F_2$ is well defined there exists an embedding ψ such that $\det(F_1^\top F_2) = \psi(F_1)^\top \psi(F_2)$. We will now show that the theorem follows by applying this result to $F_i = \Lambda^{1/2} O_i$, $i = 1, 2$, where

$$\Lambda = \begin{bmatrix} 1 & & & \\ & \lambda & & \\ & & \lambda^2 & \\ & & & \ddots \end{bmatrix} \quad O_i = \begin{bmatrix} C_i \\ C_i A_i \\ C_i A_i^2 \\ \vdots \end{bmatrix} \quad i = 1, 2. \tag{10}$$

Notice first that $F_1^\top F_2 = O_1^\top \Lambda O_2$ is well defined because

$$Q_{12} = O_1^\top \Lambda O_2 = \sum_{i=0}^{\infty} \lambda^i (A_1^\top)^i C_1^\top C_2 A_2^i \tag{11}$$

converges when $\lambda \|A_1\|_2 \|A_2\|_2 < 1$. Notice also that

$$\begin{aligned} Q_{12} &= C_1^\top C_2 + \lambda A_1^\top \Big(\sum_{i=0}^{\infty} \lambda^i (A_1^\top)^i C_1^\top C_2 A_2^i \Big) A_2 \\ &= C_1^\top C_2 + \lambda A_1^\top Q_{12} A_2. \end{aligned} \tag{12}$$

Since the solution to the Sylvester equation is unique, we have $Q_{12} = P_{12}$. Therefore, letting $\phi(\mathcal{M}_i) = \psi(\Lambda^{1/2} O_i)$, we obtain

$$\phi^\top(\mathcal{M}_1)\phi(\mathcal{M}_2) = \psi(\Lambda^{1/2} O_1)^\top \psi(\Lambda^{1/2} O_2) \tag{13}$$

$$= \det(O_1^\top \Lambda O_2) = \det(P_{12}). \tag{14}$$

Thus, $\det(P_{12})$ is a kernel, hence so is $\det(P_{12})^2$. ∎

Since k_d is a kernel, so is its normalized version

$$k'_d(\mathcal{M}_1, \mathcal{M}_2) = \frac{k_d(\mathcal{M}_1, \mathcal{M}_2)}{\sqrt{k_d(\mathcal{M}_1, \mathcal{M}_1)}\sqrt{k_d(\mathcal{M}_2, \mathcal{M}_2)}}. \tag{15}$$

The following theorem shows that this kernel is invariant with respect to a change of basis.

Theorem 2: The normalized Binet-Cauchy determinant kernel is invariant with respect to a change of basis.

The theorem follows by direct calculation:

$$\begin{aligned} k'_d(\mathcal{M}'_1, \mathcal{M}'_2) &= \frac{\det(P'_{12})^2}{\sqrt{\det(P'_{11})^2}\sqrt{\det(P'_{22})^2}} \\ &= \frac{\det(T_1^{-\top} P_{12} T_2^{-1})^2}{\sqrt{\det(T_1^{-\top} P_{11} T_1^{-1})^2}\sqrt{\det(T_2^{-\top} P_{22} T_2^{-1})^2}} \\ &= \frac{\det(P_{12})^2}{\sqrt{\det(P_{11})^2}\sqrt{\det(P_{22})^2}} = k'_d(\mathcal{M}_1, \mathcal{M}_2). \end{aligned} \tag{16}$$

The next theorem shows that the Martin kernel [1] is a particular case of the normalized Binet-Cauchy determinant kernel.

Theorem 3: When $\lambda = 1$, the normalized Binet-Cauchy determinant kernel, k'_d, coincides with the Martin kernel, k_M.

[1] Here-forth whenever we mention the determinant kernel, we refer to this initial-state-independent definition. The original Binet-Cauchy determinant kernel in [3] is computationally unwieldy and is not used in this paper.

Proof: By definition, the determinant kernel for $\lambda = 1$ is given by $k_d(\mathcal{M}_1, \mathcal{M}_2) = \det(P_{12})^2$, where P_{12} is the solution of the Sylvester equation, $P_{12} = A_1^\top P_{12} A_2 + C_1^\top C_2$. Now, following [2], the Martin kernel is the square of the product of the cosine of the subspace angles between the two LDSs, which can be computed as

$$\cos^2\theta_i = i\text{-th eigenvalue}(P_{11}^{-1}P_{12}P_{22}^{-1}P_{21}), \quad (17)$$

where P_{ij} is the solution of $P_{ij} = A_i^\top P_{ij} A_j + C_i^\top C_j$. We thus have

$$\begin{aligned} k_M(\mathcal{M}_1, \mathcal{M}_2) &= \prod_{i=1}^{n} \cos^2\theta_i = \det(P_{11}^{-1}P_{12}P_{22}^{-1}P_{21}) \\ &= \frac{\det(P_{12})^2}{|\det(P_{11})||\det(P_{22})|} \\ &= \frac{k_d(\mathcal{M}_1, \mathcal{M}_2)}{\sqrt{k_d(\mathcal{M}_1, \mathcal{M}_1)}\sqrt{k_d(\mathcal{M}_2, \mathcal{M}_2)}} \\ &= k'_d(\mathcal{M}_1, \mathcal{M}_2). \end{aligned} \quad (18)$$

■

B. From Binet-Cauchy kernels to subspace angles

In this section, we present an alternative approach to making the Binet-Cauchy kernel invariant with respect to the initial conditions. The key idea behind this new approach is to maximize the Binet-Cauchy kernel $\mathbf{x}_1^\top P_{12}\mathbf{x}_2$ with respect to $\mathbf{x}_1$ and $\mathbf{x}_2$. Interestingly, we show that different choices for the maximization lead to the cosines of the different subspace angles between LDSs.

We begin by defining the following kernel

$$\begin{aligned} k'_{tr}(\mathcal{M}_1, \mathcal{M}_2) &= \max_{\mathbf{x}_1, \mathbf{x}_2} \left(\mathbf{x}_1^\top P_{12}\mathbf{x}_2\right) \\ \text{subject to} \quad & \mathbf{x}_1^\top P_{11}\mathbf{x}_1 = 1 \text{ and } \mathbf{x}_2^\top P_{22}\mathbf{x}_2 = 1, \end{aligned} \quad (19)$$

where, as before, P_{ij} is the solution to the Sylvester equation, $P_{ij} = \lambda A_i^\top P_{ij} A_j + C_i^\top C_j$. When $\lambda = 1$, this kernel is equal to the cosine of the smallest subspace angle between two LDSs, as shown by the following result.

Theorem 4: When $\lambda = 1$, k'_{tr} coincides with the cosine of the smallest subspace angle.

Proof: The Lagrangian of the optimization problem in Equation (19) is given by:

$$\mathbf{x}_1^\top P_{12}\mathbf{x}_2 + \frac{1}{2}\mu_1(1 - \mathbf{x}_1^\top P_{11}\mathbf{x}_1) + \frac{1}{2}\mu_2(1 - \mathbf{x}_2^\top P_{22}\mathbf{x}_2).$$

Differentiating and equating to zero gives,

$$P_{12}\mathbf{x}_2 = \mu_1 P_{11}\mathbf{x}_1, \quad (20)$$

$$P_{12}^\top \mathbf{x}_1 = \mu_2 P_{22}\mathbf{x}_2. \quad (21)$$

Multiplying Equation (20) by $\mathbf{x}_1^\top$ and Equation (21) by $\mathbf{x}_2^\top$ on the right and equating them gives

$$\mu_1 \mathbf{x}_1^\top P_{11}\mathbf{x}_1 = \mathbf{x}_1^\top P_{12}\mathbf{x}_2 = \mu_2 \mathbf{x}_2^\top P_{22}\mathbf{x}_2. \quad (22)$$

Using the constraints in Equation (19), we get $\mu_1 = \mu_2 = \mu$ and thus $\mathbf{x} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top]^\top$ is the solution to the generalized eigenvalue problem:

$$\begin{bmatrix} 0 & P_{12} \\ P_{21} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mu \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}. \quad (23)$$

Following the construction in [2], μ^2 is the eigenvalue of the matrix $P_{11}^{-1}P_{12}P_{22}^{-1}P_{21}$ and $\mathbf{x} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top]^\top$ is the generalized eigenvector corresponding to the generalized eigenvalue, μ, in Equation (23). Multiplying by $\mathbf{x}^\top$ on both sides of Equation (23), we obtain $\mathbf{x}_1^\top P_{12}\mathbf{x}_2 = \mu$ and thus the solution to the optimization problem becomes

$$k'_{tr}(\mathcal{M}_1, \mathcal{M}_2) = \mu_{\max} = \cos\theta_{\min}, \quad (24)$$

which coincides with the cosine of the smallest subspace angle between systems $\mathcal{M}_1$ and $\mathcal{M}_2$. ■

More generally, the remaining subspace angles can be computed from the Binet-Cauchy trace kernels by successively solving a series of constrained optimization problems. In particular, one can show in an analogous fashion that for the k-th smallest subspace angle we have

$$\begin{aligned} &\cos\theta_k = \max_{\mathbf{x}_1, \mathbf{x}_2} \left(\mathbf{x}_1^\top P_{12}\mathbf{x}_2\right), \text{ for } k = 2, \ldots, n \\ &\text{subject to } \mathbf{x}_1^\top P_{11}\mathbf{x}_1 = 1, \mathbf{x}_2^\top P_{22}\mathbf{x}_2 = 1, \\ &\mathbf{x}_{1;i}^\top P_{11}\mathbf{x}_1 = 0, \mathbf{x}_{2;i}^\top P_{22}\mathbf{x}_2 = 0, \text{ for } i = 1, 2, \ldots, k-1. \end{aligned} \quad (25)$$

where $\mathbf{x}_{1;i}$ and $\mathbf{x}_{2;i}$ are the corresponding maximizers for $\cos\theta_i, i = 1, 2, \ldots, k-1$. Hence, the subspace angles can be directly derived from the Binet Cauchy kernels with $\lambda = 1$ and therefore, the subspace-angle based distances are special cases of the Binet Cauchy kernels.

IV. Hybrid metrics on the output means and system dynamics

An important consideration that is often overlooked when developing metrics is how to incorporate the effect of the temporal means when computing the distances. It is clear that the temporal means of two sequences provide good discriminative power for recognition purposes. Using the temporal means alone as weak classifiers with Boosting has been shown to perform well in [36].

We can define a simple metric that uses only the temporal means of the output sequences:

$$d_p(\mu_1, \mu_2) = \|\mu_1 - \mu_2\|^p, \quad (26)$$

where $p \geq 1$ is a free parameter, usually equal to 1.

The distances based on subspace angles, Binet-Cauchy kernels and in general any dynamical system metric can be combined with this metric on the temporal means to construct a new class of *hybrid metrics* that also give a certain weight to the temporal means when performing recognition. This class of hybrid distances can in general be represented by:

$$d_h(\mathcal{M}_1, \mathcal{M}_2) = (1 - \beta)d_c(\mathcal{M}_1, \mathcal{M}_2) + \beta d_p(\mathcal{M}_1, \mathcal{M}_2), \quad (27)$$

where d_c is any metric between the LDSs and d_p is the distance between the temporal means. Note that d_c and d_p are normalized and scaled such that the maximum distance between any two models in the training set is one. The parameter β is the relative weight between d_c and d_p and can be tuned using cross-validation. Also, notice that for all the metrics in §II, the distances can easily be converted into Radial Basis Function (RBF) kernels with a parameter γ as

$k(\mathcal{M}_1, \mathcal{M}_2) = \mathbf{e}^{-\gamma d(\mathcal{M}_1, \mathcal{M}_2)^2}$. This conversion allows γ to be tuned to the specific application using cross-validation during the training phase.

V. Experiments

In this section, we will provide experimental results for human activity recognition and compare the performance of using our proposed initial-state invariant Binet-Cauchy kernel against the original Binet-Cauchy kernel as well as the more commonly used Martin distance. In the following, we will first briefly describe the feature extraction procedure and various parameter choices, and then provide results on several human activity databases.

A. Feature extraction

We use the Histograms of Oriented Optical Flow (HOOF) features proposed by Chaudhry et al. [18] since they do not require any pre-processing of the video such as human tracking, background subtraction, or silhouette extraction as long as there is only one person in the scene and the camera is stationary. HOOF features are extracted from each frame by first computing the optical flow of each frame, quantizing the flow directions in a number of laterally invariant bins[2] and adding the magnitude of the flow vector at each pixel to the corresponding bin before normalizing the histogram. This results in a feature that is invariant to the lateral direction of motion (a person moving left to right vs right to left will generate the same signature), and invariant to scale (a person further away in the scene will generate similar optical flow signatures as a person who is near the camera).

Parameter choices. There are several parameter choices when using HOOF features and kernel NLDS for representing human actions, including the number of histogram bins, B, and the choice of the histogram kernel, e.g., Geodesic, χ^2, Minimum Distance Pairwise Assignment (MDPA) or Histogram Intersection (HIST). The order of the dynamical system, n, is another parameter, as is the choice of using dynamics-only metrics or hybrid metrics, as discussed in §IV. From a preliminary set of experiments on the Weizmann human action dataset [37], we found that in general, any bin size, $B > 20$ is discriminative across all metrics and histogram kernels. Furthermore, we found that overall, lower system orders, $n \approx 5$, the geodesic kernel for histograms and the histogram intersection kernel performed better.

B. Experiments on the Weizmann database [37]

The Weizmann human action dataset consists of a total of 93 videos with 9 actors and 10 action categories including both in-place actions such as waving, bending, etc. , and moving actions such as walking, running, etc. . The commonly used testing scheme is a leave-one-sequence out cross-validation approach and Nearest-Neighbor classification.

We will first provide several statistics of how the recognition performance varies across different dynamical systems-based metrics and histogram kernels chosen for HOOF. We will use *boxplots* to show these statistics for each choice of metric and histogram kernel for a range of bin sizes (20-100) and system orders (1-20). Figure 1 shows a generic box-plot.

[2]An optical flow vector, (x, y) contributes to the same bin as $(-x, y)$.

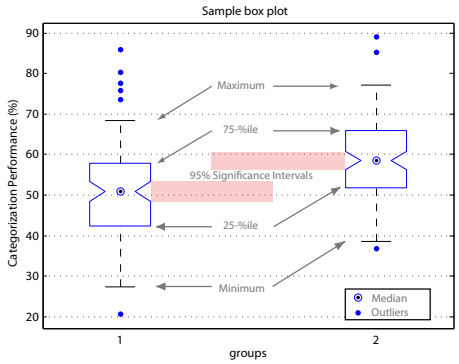

Fig. 1. Example of a box plot used to compare the statistics of two sets of categorization performances. The graphs display the non-outlier minimum and maximum performance, the 25- and 75- percentiles, median and outliers. The notches around the median represent ranges for statistical significance tests at 95% for difference in median performance.

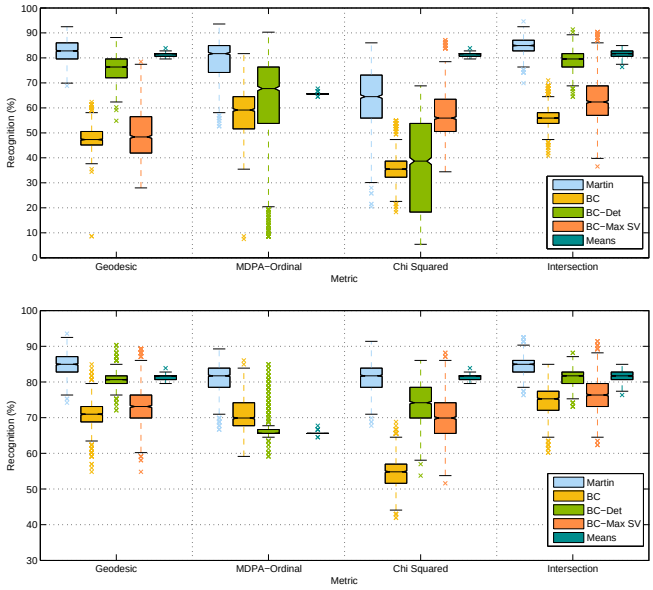

Fig. 2. Performance statistics for different kernels and LDS metrics on the Weizmann database. Top: Dynamics-only metrics, Bottom: Hybrid metrics. (Note that for the MDPA kernel the BC-Max SV results are missing. This is because we cannot compute an explicit embedding for HOOF time-series and hence we cannot use LDS system parameters and therefore the JCF that is required to compute the Binet-Cauchy maximum singular value kernel.)

Figure 2 shows the statistics of recognition performance against several histogram metrics used for system identification and dynamical systems metrics. We can see that in general, the Geodesic metric (or Bhattacharyya kernel) and the Histogram Intersection kernel (HIST) between two histograms give the best results. Furthermore, the best recognition results are achieved when using the Martin distance followed by the Binet-Cauchy initial state independent determinant kernel. The median performance of the Binet-Cauchy determinant kernel is better than the Binet-Cauchy maximum singular value kernel and significantly better than

TABLE I

COMPARISON OF DIFFERENT APPROACHES FOR ACTION RECOGNITION ON THE WEIZMANN DATABASE AGAINST OUR PROPOSED METHODS.

Method	Recognition (%)
Xie et al. [38]	95.60
Thurau et al. [39]	94.40
Ikizler et al. [40]	100.00
Gorelick et al. [37]	99.60
Niebles et al. [41]	90.00
Ali et al. [42]	95.75
HOOF - Martin distance	94.62
HOOF - Binet-Cauchy kernel	84.95
HOOF - Binet-Cauchy determinant kernel	92.47

the original Binet-Cauchy kernel. This shows the importance of having an initial-state invariant metrics as the best results are achieved using the Martin distance and the Binet-Cauchy determinant kernel which are both invariant w.r.t. initial states of the dynamical systems.

Table I compares the performance of state-of-the-art methods on the Weizmann dataset and dynamical-systems based approaches using the Martin distance, the original Binet-Cauchy kernel and our proposed initial-state invariant Binet-Cauchy determinant kernel. As we can see, using the Binet-Cauchy determinant kernel gives comparable results with several other methods and performs much better than the original Binet-Cauchy kernel.

C. More datasets

Given the above insights, we will now test the performance of our proposed metric on some other human activity datasets.

Multi-view human action dataset. We collected a high-resolution multiple view dataset of 12 subjects performing 11 actions including jumping, sitting, throwing, etc. , with 5 repetitions of each action. Table II provides activity recognition results independently for each of the four different view points. The results shown are the best across orders 1-20 when using hybrid metrics with a bin size of 64 and the Geodesic kernel. Overall the best results are achieved when using SVM coupled with either the Martin distance or the Binet-Cauchy determinant kernel. Table III shows the best results achieved using the common bag-of-words approach with HOG-HOF features as in [43] along-with the χ^2 kernel for bag-of-words histograms. Note that our results when using HOOF are competitive with these results and as noted before, the Binet-Cauchy determinant kernel performs much better than the original Binet-Cauchy kernel.

KTH dataset [44]. We also provide classification results using HOOF on the KTH database. This is a very challenging dataset for global features such as HOOF since there is a lot of camera jitter and camera artifacts such as automatic white balance and exposure adjustment, etc. These artifacts cause errors in optical flow computation which affect the accuracy of HOOF computation. Furthermore, there are several frames at the beginning and end of a video that do not contain a person. Therefore we only provide results for scenario 1 of

TABLE II

ACTION CLASSIFICATION RESULTS ON THE COLLECTED MULTI-VIEW ACTION DATASET USING HOOF FROM DIFFERENT CAMERA VIEWS (V-1 THROUGH V-4).

	Metric	V-1	V-2	V-3	V-4
1-NN	Martin	80.00	80.36	88.69	76.28
	BC	64.36	70.18	77.01	71.53
	BC-Det	**81.82**	**83.64**	**89.78**	**79.20**
	BC-Max SV	75.27	83.64	85.04	74.82
	Means	65.09	72.00	75.91	64.60
SVM	Martin	**87.27**	**89.09**	93.07	**86.86**
	BC	70.18	75.64	87.23	79.20
	BC-Det	85.82	84.72	**94.89**	84.31
	BC-Max SV	80.73	82.18	89.05	82.12
	Means	61.45	57.09	85.04	74.45

TABLE III

ACTION CLASSIFICATION RESULTS ON THE COLLECTED MULTI-VIEW ACTION DATASET USING HOG-HOF FROM DIFFERENT CAMERA VIEWS (V-1 THROUGH V-4).

	V-1	V-2	V-3	V-4
1-NN	91.61	91.97	93.43	84.67
SVM	91.97	87.96	96.35	87.23

the KTH database. Table IV provides activity recognition results for several dynamical-systems based metrics when using the Geodesic kernel for 64-bin HOOF time-series with a system order of 5. The best recognition rate achieved was 72.56% with the Martin kernel using SVM followed closely by the Binet-Cauchy determinant kernel, which again performed much better than the original Binet-Cauchy kernel.

We would like to note that even though the KTH dataset is one of the most commonly evaluated upon dataset in computer vision, almost all state-of-the-art approaches are based on local features and hence are not directly comparable to our approach. To the best of our knowledge the best performing global feature-based method for KTH was outlined in [45] as a comparison method against their proposed local feature-based methods. The average recognition rate achieved by this method was 72%. Another recent global optical-flow feature based approach was proposed by Mota et al. [46] that gave recognition rates in the range of 70% to 86% on the KTH dataset. Unfortunately, all global-feature based representations do not fare well against the best reported local-feature based result on the entire (all four scenarios) KTH database, e.g., 98.1% in [47].

UCF50 dataset [48]. Finally, we will also report activity recognition results for the large 50-class UCF50 dataset. Table V shows the recognition rates for several metrics when using 64-bin HOOF time-series modeled using the Geodesic kernel dynamical systems. We report the average of 5-fold group-wise classification results as in [49], [47]. The results in Table V follow the trends that we have observed for other datasets: in general hybrid metrics perform better than dynamics-only metrics, and Martin and Binet-Cauchy determinant kernel perform the best. The best performing result of 53.14% is achieved by using the Binet-Cauchy

TABLE IV

ACTIVITY RECOGNITION RESULTS ON THE KTH DATABASE. HOOF RESULTS ONLY CORRESPOND TO SCENARIO 1 USING THE COMMON 16/9 SUBJECT TRAIN/TEST SPLIT. WE USED THE GEODESIC KERNEL, 64 BINS AND SYSTEM ORDER 5 FOR MODELING HOOF TIME-SERIES. THE RESULTS FOR STATE-OF-THE-ART GLOBAL AND LOCAL METHODS ARE MARKED WITH AN ASTERISK AS THE RESULTS ARE NOT DIRECTLY COMPARABLE DUE TO DIFFERENT TRAINING/TESTING SETS OR DIFFERENT NUMBER OF SCENARIOS.

Method/Distance	Dynamics-only		Hybrid	
HOOF	1-NN	SVM	1-NN	SVM
Martin	66.51	**72.56**	**67.44**	**67.91**
BC	49.30	46.51	59.07	62.79
BC-Det	**66.98**	61.40	60.93	60.47
BC-Max SV	27.91	29.77	48.84	41.86
Means			56.74	61.86
Laptev et al. [45]	72*			
Mota et al. [46]	70-86*			
Sadanand et al. [47]	98.1*			

TABLE V

ACTIVITY RECOGNITION RESULTS FOR THE UCF50 DATABASE USING 5-FOLD GROUP-WISE CROSS-VALIDATION. FOR OUR METHOD, WE USED 64 BINS, THE GEODESIC KERNEL AND SYSTEM ORDER 5 FOR MODELING HOOF TIME-SERIES. THE STATE-OF-THE-ART RESULTS USING THE METHOD IN [50] AND [43] APPEARED IN [47].

Method/Distance	Dynamics-only		Hybrid	
HOOF	1-NN	SVM	1-NN	SVM
Martin	41.62	26.19	**46.33**	34.43
BC	25.43	31.77	32.08	42.14
BC-Det	**42.00**	**48.51**	45.71	**53.14**
BC-Max SV	17.19	18.62	24.30	31.94
Means			31.67	33.15
Oliva et al. [50]	38.8			
Laptev et al. [43]	47.9			
Sadanand et al. [47]	57.9			

determinant kernel. The UCF dataset is relatively new and has only recently started to gain the attention of researchers. The best reported results using simple global appearance-based features such as Gist [50] is 38.8%, using local spatio-temporal HOG-HOF feature-based bag-of-words approach [43] is 47.9%, and using the approach in [47] is 57.9%. Our result of 53.14% is better than two of these approaches and is competitive with the state-of-the-art.

VI. CONCLUSIONS

In this paper, we have proposed initial-state invariant versions of the Binet-Cauchy kernel. We have shown that our proposed kernels are theoretically sound and that they allow us to develop interesting connections between the Binet-Cauchy kernels and the more commonly used Martin and subspace angle-based distances for dynamical systems. Through our experiments, we have shown that the initial-state invariant kernels perform much better than the original Binet-Cauchy kernels for the task of activity recognition.

REFERENCES

[1] A. Martin, "A metric for ARMA processes," *IEEE Trans. on Signal Processing*, vol. 48, no. 4, pp. 1164–1170, 2000.

[2] K. D. Cock and B. D. Moor, "Subspace angles and distances between ARMA models," *System and Control Letters*, vol. 46, no. 4, pp. 265–270, 2002.

[3] S. Vishwanathan, A. Smola, and R. Vidal, "Binet-Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes," *International Journal of Computer Vision*, vol. 73, no. 1, pp. 95–119, 2007.

[4] B. Béjar, L. Zappella, and R. Vidal, "Surgical gesture classification from video data," in *Medical Image Computing and Computer Assisted Intervention*, 2012, pp. 34–41.

[5] A. Ravichandran, R. Vidal, and H. Halperin, "Segmenting a beating heart using polysegment and spatial GPCA," in *IEEE International Symposium on Biomedical Imaging*, 2006, pp. 634–637.

[6] A. Ghoreyshi and R. Vidal, "Epicardial segmentation in dynamic cardiac MR sequences using priors on shape, intensity, and dynamics, in a level set framework," in *IEEE International Symposium on Biomedical Imaging*, 2007, pp. 860–863.

[7] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[8] L. Yuan, F. Wen, C. Liu, and H. Shum, "Synthesizing dynamic texture with closed-loop linear dynamic system," in *European Conference on Computer Vision*, 2004, pp. 603–616.

[9] R. Vidal and A. Ravichandran, "Optical flow estimation and segmentation of multiple moving dynamic textures," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. II, 2005, pp. 516–521.

[10] A. Ravichandran and R. Vidal, "Video registration using dynamic textures," in *European Conference on Computer Vision*, 2008.

[11] A. Ravichandran, R. Chaudhry, and R. Vidal, "View-invariant dynamic texture recognition using a bag of dynamical systems," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[12] A. Ravichandran and R. Vidal, "Video registration using dynamic textures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 158–171, January 2011.

[13] A. Ravichandran, R. Chaudhry, and R. Vidal, "Categorizing dynamic textures using a bag of dynamical systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[14] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto, "Recognition of human gaits," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2001, pp. 52–58.

[15] A. Bissacco, A. Chiuso, and S. Soatto, "Classification and recognition of dynamical models: The role of phase, independent components, kernels and optimal transport," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1958–1972, 2007.

[16] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[17] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition," in *IEEE International Conference on Computer Vision*, 2007.

[18] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[19] B. Li, M. Ayazoglu, T. Mao, O. I. Camps, and M. Sznaier, "Activity recognition using dynamic subspace angles," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3193–3200.

[20] G. Aggarwal, A. Roy-Chowdhury, and R. Chellappa, "A system identification approach for video-based face recognition," in *IEEE International Conference on Pattern Recognition*, 2004, pp. 23–26.

[21] P. Saisan, A. Bissacco, A. Chiuso, and S. Soatto, "Modeling and synthesis of facial motion driven by speech," in *European Conference on Computer Vision*, vol. 3, 2004, pp. 456–467.

[22] P. S. Krishnaprasad, "Geometry of minimal systems and the identification problem," Ph.D. dissertation, Harvard University, 1977.

[23] P. S. Krishnaprasad and C. F. Martin, "On families of systems and deformations," *International Journal of Control*, vol. 38, no. 5, pp. 1055–l079, 1983.

[24] B. Hanzon and S. I. Marcus, "Riemannian metrics on spaces of stable linear systems, with applications to identification," in *IEEE Conference on Decision & Control*, 1982, pp. 1119–1124.

[25] B. Hanzon, *Identifiability, Recursive Identification and Spaces of*

Linear Dynamical Systems. Centrum voor Wiskunde en Informatica (CWI), 1989, vol. 63-64.

[26] R. M. Gray, *Probability, random processes, and ergodic properties*. Springer, 2009.

[27] S. I. Amari, "Differential geometry of a parametric family of invertible linear systems-Riemannian metric, dual affine connections, and divergence," *Mathematical Systems Theory*, vol. 20, pp. 53–82, 1987.

[28] S. I. Amari and H. Nagaoka, *Methods of Information Geometry*, ser. Translations of Mathematical Monographs. American Mathematical Society, 2000, vol. 191.

[29] T. T. Georgiou, "Distances and Riemannian metrics for spectral density functions," *IEEE Transactions on Signal Processing*, vol. 55, no. 8, pp. 3995–4003, August 2007.

[30] X. Jiang, L. Ning, and T. T. Georgiou, "Distances and Riemannian metrics for multivariate spectral densities," *IEEE Transactions on Automatic Control*, vol. 57, no. 7, pp. 1723–1735, 2012.

[31] B. Afsari, R. Chaudhry, A. Ravichandran, and R. Vidal, "Group action induced distances for averaging and clustering linear dynamical systems with applications to the analysis of dynamic visual scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[32] A. Chan and N. Vasconcelos, "Probabilistic kernels for the classification of auto-regressive visual processes," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 846–851.

[33] P. V. Overschee and B. D. Moor, "N4SID : Subspace algorithms for the identification of combined deterministic-stochastic systems," *Automatica, Special Issue in Statistical Signal Processing and Control*, pp. 75–93, 1994.

[34] R. Shumway and D. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[35] A. Chan and N. Vasconcelos, "Classifying video with kernel dynamic textures," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[36] R. Vidal and P. Favaro, "Dynamicboost: Boosting time series generated by dynamical systems," in *IEEE International Conference on Computer Vision*, 2007.

[37] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[38] Y. Xie, H. Chang, Z. Li, L. Liang, X. Chen, and D. Zhao, "A unified framework for locating and recognizing human actions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[39] C. Thurau and V. Hlavac, "Pose primitive based human action recognition in videos or still images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[40] N. Ikizler and P. Duygulu, "Histogram of oriented rectangles: A new pose descriptor for human action recognition," *Image and Vision Computing*, vol. 27 (10), pp. 1515–1526, 2009.

[41] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, pp. 299–318, 2008.

[42] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32 (2), pp. 288–303, 2010.

[43] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[44] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *IEEE International Conference on Pattern Recognition*, 2004.

[45] I. Laptev, B. Caputo, C. Schüldt, and T. Lindberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Computer Vision and Image Understanding*, vol. 108, pp. 207–229, 2007.

[46] V. F. Mota, E. A. Perez, M. B. Vieira, L. M. Maciel, F. Precioso, and P. H. Gosselin, "A tensor based on optical flow for global description of motion in videos," in *Proceedings of Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2012.

[47] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[48] "UCF50 dataset." [Online]. Available: http://vision.eecs.ucf.edu/data.html

[49] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *IEEE International Conference on Computer Vision*, 2011.

[50] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42 (3), pp. 145–175, 2001.