International Journal of Computer Vision
DOI 10.1007/s11263-016-0918-1

Complex Activity Recognition via Attribute Dynamics

Wei-Xin Li · Nuno Vasconcelos

Received: 9 July 2015 / Accepted: 26 May 2016

Abstract The problem of modeling the dynamic structure of human activities is considered. Video is mapped to a semantic feature space, which encodes activity attribute probabilities over time. The binary dynamic system (BDS) model is proposed to jointly learn the distribution and dynamics of activities in this space. This is a non-linear dynamic system that combines binary observation variables and a hidden Gauss-Markov state process, extending both binary principal component analysis (PCA) and the classical linear dynamic systems (LDS). A BDS learning algorithm, inspired by the popular dynamic texture, and a dissimilarity measure between BDSs, which generalizes the Binet-Cauchy kernel, are introduced. To enable the recognition of highly non-stationary activities, the BDS is embedded in a bag of words. An algorithm is introduced for learning a BDS codebook, enabling the use of the BDS as a visual word for attribute dynamics (WAD). Short-term video segments are then quantized with a WAD codebook, allowing the representation of video as a bag-of-words for attribute dynamics (BoWAD). Video sequences are finally encoded as vectors of locally aggregated descriptors (VLAD), which summarize the first-moments of video snippets on the BDS manifold. Experiments show that this representation achieves state-of-the-art performance on the tasks of complex activity recognition and event identification.

Keywords Complex activity · Attribute · Dynamical model · Variational inference · Fisher score

W.-X. Li
E-mail: wel017@ucsd.edu

N. Vasconcelos
E-mail: nuno@ece.ucsd.edu

ECE Department, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093, United States

1 Introduction

Understanding human behavior is an important goal for computer vision (Aggarwal and Ryoo, 2011). While early solutions mostly addressed the recognition of simple behavior in controlled environments (Bregler, 1997; Bobick and Davis, 2001; Schuldt et al, 2004; Gorelick et al, 2007), recent interest has been in more challenging and realistic tasks (Laptev et al, 2008; Rodriguez et al, 2008; Niebles et al, 2010; Kuehne et al, 2011). In the literature, these tasks are commonly referred to as "action" or "activity" recognition. In this work, we adopt the term "action" to denote movements at the lowest level of the semantic hierarchy, *e.g.*, "run," "jump," or "kick a ball". The term "activity" is reserved for behavior of higher level semantics, which can usually be described as a sequence of actions. For example, the Olympic activity "clean and jerk" involves the actions of "grasping a barbell," "raising weights over the athlete's head," and "dropping the bar". Activities can also be performed by multiple subjects (*i.e.*, be "collective"), or composed of "events" rather than actions (*e.g.*, "wedding ceremony" composed of events such as "walking the bride," "exchange of vows," "opening dance," *etc*).

Several of the prior works in action and activity recognition have proposed variants of the *bag of visual words* (BoVW), which represents video as a collection of orderless spatiotemporal features and serves as the low-level foundation for many other action analysis frameworks. This family of representations have been shown to consistently achieve state-of-the-art performance for tasks such as action recognition and retrieval (Wang et al, 2009; Tamrakar et al, 2012; Wang and Schmid, 2013; Peng et al, 2014; Ni et al, 2015; Lan et al, 2015).

Published online: 21 June 2016

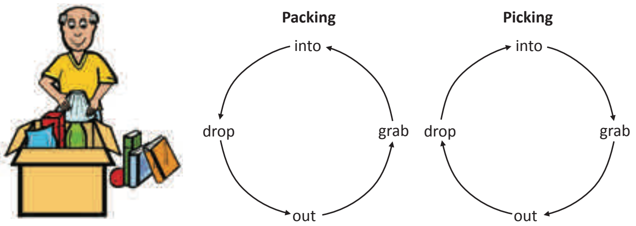


Fig. 1: The actions "move hand into box" (into), "grab object" (grab), "move hand out of box" (out), and "drop object" (drop) are consisting with the activities of "packing a box" and "picking objects from a box". In the absence of temporal modeling of event semantics, these activities can be quite difficult to distinguish.

Nevertheless, the BoVW has at least two important limitations. First, it does not account for the fact that most activities are best abstracted as sequences of actions or events. This is illustrated by the activity "packing a box" of Figure 1, which most humans would characterize as a sequence of the actions "move hand out of box - grab object - move hand into box - drop object." In the absence of an explicit representation of these semantics, it is up to the classifier to learn the importance of concepts such as moving hands, grabbing or dropping objects for the characterization of this activity. While these concepts are not impossible to learn from the evolution of low-level features, this is easier when the classifier is given explicit supervision about the semantics of interest. In result, semantic video modeling has recently began to receive substantial attention. For example, the TRECVID multimedia event *detection* and *recounting* contest (Over et al, 2011), one of the major large-scale video analysis research efforts, explicitly states the goal of not only predicting the event category ("detection") of a video sequence, but also identifying its *semantically meaningful and relevant pieces* ("recounting").

Second, the BoVW captures little information about the temporal structure of video. This limits its expressiveness, since a single set of actions (or events) can give rise to multiple activities, depending on the *order* with which the actions are performed. This is again illustrated in Figure 1, where the activity of "picking objects from a box" differs from the activity of "packing a box" only in terms of the order of the actions described above, which is now "move hand into box - grab object - move hand out of box - drop object". Hence, sophisticated modeling of temporal structure can be critical for parsing complex activities. This is beyond the reach of the BoVW.

Recently, there have been various attempts to address the two limitations of the BoVW. On one hand, several authors have proposed richer models of the temporal structure, also known as *dynamics*, of human activity (Niebles et al, 2010; Laxton et al, 2007; Chaudhry et al, 2009; Gaidon et al, 2011). However, because modeling activity dynamics can be a complex proposition, it is not uncommon for these models to require features specific to certain data sets or activity classes (Laxton et al, 2007; Chaudhry et al, 2009), or non-trivial forms of pre-processing, such as tracking (Li et al, 2011), per-class manual annotation (Gaidon et al, 2011), *etc.* On the other hand, inspired by recent developments in image classification (Lampert et al, 2009; Rasiwasia and Vasconcelos, 2012), there has been a move towards the representation of action in terms of intermediate-level semantic concepts, such as *attributes* (Liu et al, 2011; Fathi and Mori, 2008). This introduces a layer of abstraction that improves generalization, enables modeling of contextual relationships (Rasiwasia and Vasconcelos, 2009), and simplifies knowledge transfer across activity classes (Liu et al, 2011). However, these models continue to disregard the temporal structure of video.

In this work, we propose an activity representation that combines all these properties, by *modeling the dynamics of human activities in the space of attributes*. The idea is to define each activity as a sequence of *semantic* events, *e.g.*, defining "packing a box" as the *sequence* of the action attributes "remove (hand from box)", "grab (object)", "insert (hand in box)", and "drop (object)". This semantic-level representation is *more robust* to confounding factors, such as diversity of grabbing styles, hand motion speeds, or camera motion, than dynamic representations based on low-level features. It is also *more discriminant* than semantic representations that ignore dynamics, *i.e.*, that simply record the occurrence (or frequency) of the action attributes "remove", "grab", "insert", and "drop". We already saw that, in the absence of information about the *sequence* in which these attributes occur, the "packing a box" activity cannot be distinguished from the "picking from a box" activity.

To implement this idea, we present novel solutions to the two major technical challenges of using attribute dynamics for activity recognition. The first is the modeling of attribute dynamics itself. As usual in semantics-based recognition (Liu et al, 2011), video is represented in a semantic feature space, where each feature encodes the probability of occurrence of an action attribute at each time step. We introduce a generative model, the *binary dynamic system* (BDS), to learn *both* the distribution and dynamics of different activities in this space. The BDS is a non-linear dynamic system that combines binary observations with a hidden Gauss-Markov state process. It can be interpreted as either 1) a generalization of *binary principal component analysis* (binary PCA) (Schein et al, 2003), which accounts for data dynamics; or 2) an extension of the classical *linear dynamic system* (LDS) to a binary observation space.

The second is to account for non-stationary video dynamics. For this, we embed the BDS in the BoVW representation, modeling video sequences as orderless combinations of *short-term video segments of characteristic semantic dynamics*. More precisely, videos are modeled as sequences of short-term segments sampled from a family of BDSs. This representation, the *bag of words for attribute dynamics* (BoWAD), is applicable to more complex activities, *e.g.*, "moving objects across two boxes" which combines the event sequences of "picking objects from a box" and "packing a box," with potentially other events (*e.g.*, "inspecting object") in between. The BoWAD is shown to cope with the semantic noise, content irregularities, and intra-class variation that prevail in video of complex activities.

Various tools are introduced to perform learning and inference with both the BDS and the BoWAD. We start with an efficient procedure for learning BDS parameters, inspired by the popular *dynamic texture* of (Doretto et al, 2003), combining binary PCA and a least squares problem. We then derive a dissimilarity measure between BDSs, which generalizes the Binet-Cauchy kernel from the LDS literature (Vishwanathan et al, 2006) and is used to find the nearest-neighbors of a BDS. Finally, a novel clustering algorithm, explicitly designed to cluster attribute sequences in the BDS domain, is proposed to learn the BDS codebook at the core of the BoWAD representation.

These learning tools are complemented by a discriminating feature representation for activity classification, inspired by the recent success of Fisher vectors in image classification (Perronnin et al, 2010; Krapac et al, 2011; Cinbis et al, 2012; Simonyan et al, 2013). While the BoWAD encodes zeroth moments of the cluster assignments of a video sequence to a BDS codebook, Fisher vectors complement these with first and second (central) moments statistics. This improves discrimination but has higher complexity. The *vector of locally aggregated descriptors* (VLAD) of (Jegou et al, 2012), based only on first moments, has most of the advantages of the Fisher vector but substantially less computation. We extend the VLAD to the BoWAD by introducing the *vector of locally aggregated descriptors for attribute dynamics* (VLADAD). This turns out to be computationally intractable, but can be approximated with resort to variational inference techniques (Jordan et al, 1999). We derive an implementation of the VLADAD based on the aggregation of the derivatives of a variational lower-bound of the log-likelihood over attribute sequences. Experimentally, the combination of the VLADAD with a linear classifier outperforms recent approaches to activity recognition based on dynamics and attributes.

Preliminary versions of this work were presented in (Li and Vasconcelos, 2012; Li et al, 2013b). A preliminary discussion of the BDS was presented in (Li and Vasconcelos, 2012) and a preliminary discussion of the BoWAD in (Li et al, 2013b). In addition to a unified development of these representations, the current paper introduces a number of extensions. These include 1) a novel interpretation of the VLAD as an encoding scheme in the model manifold, which enables its application to dynamic systems; 2) a new variational framework for BDS inference, which addresses the intractability of exact inference with this model; 3) the use of this framework to derive the VLAD descriptor associated with the BDS model; 4) a recounting procedure for the identification of video segments informative of target activities; and 5) an extensive empirical study of the performance of the VLADAD descriptor, involving new baselines, larger benchmarks, updated state of the art results, and a more in-depth analysis.

2 Related Work

Many approaches to action recognition have been proposed in the last decades (Aggarwal and Ryoo, 2011; Vrigkas et al, 2015). Early methods aimed to detect a small number of short-term atomic movements in distractor-free environments. These methods relied extensively on operations such as tracking (Niyogi and Adelson, 1994; Campbell and Bobick, 1995; Moore et al, 1999), or filtering (Bregler, 1997; Pinhanez and Bobick, 1998; Yacoob and Black, 1998; Chomat and Crowley, 1999), that do not generalize well to more complex environments.

Over the last decade, there has been an increased focus on effective and scalable automatic analysis of video involving complicated motion, distractor-ridden scenes, complex backgrounds, unconstrained camera motion, *etc.* Various representations have been proposed to address these challenges, including BoVW (Schuldt et al, 2004; Laptev, 2005), spatio-temporal pyramid matching (Laptev et al, 2008; Lan et al, 2014), decomposable segments (Niebles et al, 2010; Gaidon et al, 2013), trajectories (Matikainen et al, 2010; Jiang et al, 2012; Wang et al, 2013; Wang and Schmid, 2013), attributes (Liu et al, 2011), fusion with depth-maps (Yu et al, 2015), holistic volume encoding (Gorelick et al, 2007; Rodriguez et al, 2008; Shao et al, 2014), neural networks (Ji et al, 2013; Simonyan and Zisserman, 2014; Ng et al, 2015; Wang et al, 2015), and so forth. In this context, the

BoVW and its variants have consistently achieved state-of-the-art performance for tasks like action recognition and retrieval, specially when combined with informative descriptors (Laptev, 2005; Wang et al, 2009; Kovashka and Grauman, 2010; Wang and Schmid, 2013) and advanced encoding schemes (Laptev et al, 2008; Tamrakar et al, 2012; Peng et al, 2016; Shao et al, 2015). In fact, even sophisticated deep learning models, which capture hierarchical structure and have obliterated the performance of the state of the art in areas such as image and speech analysis (Deng and Yu, 2014; Russakovsky et al, 2015; Szegedy et al, 2015), have failed to match the most recent BoVW schemes based on hand-crafted features (Peng et al, 2016, 2014; Ni et al, 2015; Lan et al, 2015), in the context of action recognition from video (Simonyan and Zisserman, 2014; Ng et al, 2015; Wang et al, 2015). [1]

The main justification for the robustness of the BoVW, *i.e.*, that it reduces video to an orderless collection of spatiotemporal descriptors, also limits the applicability of this representation to fine-grained activity discrimination, where it is important to account for precise temporal structure. A number of approaches have been proposed to characterize this structure. One possibility is to represent activities in terms of limb or torso motion, spatiotemporal shape models, or motion templates (Gorelick et al, 2007; Ikizler and Forsyth, 2008). Since they require detection, segmentation, tracking, or 3D structure recovery of body parts, these representations can be fragile.

A more robust alternative is to model the temporal structure of the BoVW. This can be achieved with generalizations of popular still image recognition methods. For example, Laptev et al (2008) extend pyramid matching to video, using a 3D binning scheme that roughly characterizes the spatio-temporal structure of video. Niebles et al (2010) employ a latent support vector machine (SVM) that augments the BoVW with temporal context, which they show to be critical for understanding realistic motion. These approaches have relatively coarse modeling of dynamics. More elaborate models are usually based on generative representations. For example, Laxton et al (2007) model a combination of object contexts and motion sequences with a dynamic Bayesian network, while Gaidon et al (2011) reduce each activity to three atomic actions and model their temporal distributions. These methods rely on activity-class specific features and require detailed manual supervision. Alternatively, several researchers have proposed to model BoVW dynamics with LDSs. For example, Kellokumpu et al (2008) combine dynamic textures (Doretto et al, 2003) and local binary patterns, Li et al (2011) perform a discriminant canonical correlation analysis on the space of activity dynamics, and Chaudhry et al (2009) map frame-wise motion histograms to a reproducing kernel Hilbert space, where they learn a kernel dynamic system (KDS).

Due to their success in areas like handwriting (Graves and Schmidhuber, 2009) and speech recognition (Graves et al, 2013), *recurrent neural networks* (RNN) have recently started to receive substantial attention for activity recognition. In this context, they are usually learned from features extracted with a low-level visual representation (BoVW, CNN, *etc*). For example, Baccouche et al (2010) use an RNN to learn temporal dynamics of either hand-drafted, or CNN (Baccouche et al, 2011) features. More recently, Donahue et al (2015) combine a CNN and the *long short-term memory* (LSTM) model of (Hochreiter and Schmidhuber, 1997) to optimize both the low-level visual activation and dynamic components of an action recognition system. Alternatively, Ng et al (2015) study temporal aggregation strategies for video classification by either pooling over time or using LSTMs over frame-wise CNN activations. So far, RNN-based methods for action recognition have failed to outperform even approaches without temporal order modeling *e.g.*, the convolutional pooling of (Ng et al, 2015) or the two stream method of (Simonyan and Zisserman, 2014). A major obstacle to these approaches is temporal scalability. Since the temporal depth of a RNN is linear in the number of input frames, most methods operate on a small number of video frames, *e.g.*, 9 frames in (Baccouche et al, 2010), a few seconds in (Baccouche et al, 2011), 16 and 30 frames for (Donahue et al, 2015) and (Ng et al, 2015), respectively. This limits discrimination for complex, longer-term, activities. Finally, current RNNs model the entire content of a video sequence. This is problematic when the video contains sub-regions that do not depict the specific activity of interest, a common occurrence for open-source videos of complex activities.

Recent research in image recognition has shown that various limitations of the BoVW are overcome by representations of higher semantic level (Rasiwasia and Vasconcelos, 2012). The features that underly these representations are confidence scores for the appearance of

[1] There is an ongoing debate on how deep architectures can capture long-term low-level motion information. While early models failed to achieve competitive performance (Ji et al, 2013; Karpathy et al, 2014), recent works (Simonyan and Zisserman, 2014; Ng et al, 2015; Wang et al, 2015) show promising results, albeit still inferior to those of the best hand-crafted features (Peng et al, 2014; Ni et al, 2015; Lan et al, 2015; Peng et al, 2016). It is worth noting that this issue is orthogonal to the contributions of this work, since the proposed method is built on a space of attribute responses which could be computed with a convolutional neural network (CNN).

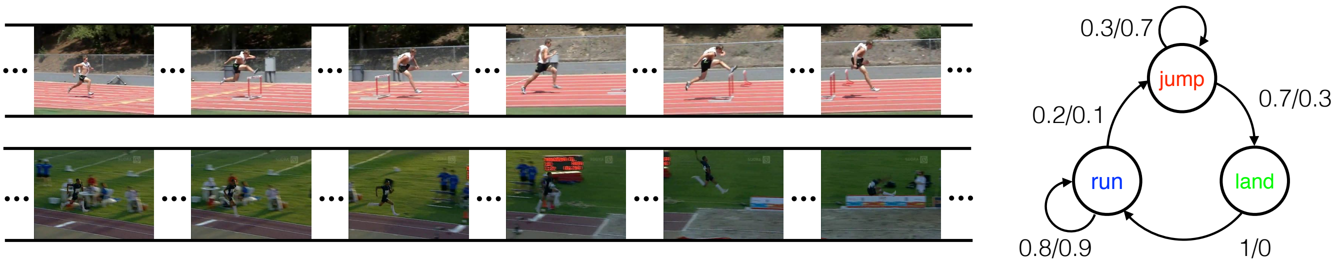


Fig. 2: Left: key frames of activities "hurdle race" (top) and "long jump" (bottom); Right: attribute transition probabilities of the two activities ("hurdle race" / "long jump") for attributes "run", "jump", and "land".

pre-defined visual concepts in images. These can be object attributes (Lampert et al, 2009), object classes (Rasiwasia and Vasconcelos, 2008; Quattoni et al, 2007; Jain et al, 2015), contextual classes (Rasiwasia and Vasconcelos, 2009), or generic visual concepts (Rasiwasi et al, 2007). Lately, semantic attributes have been used for action recognition (Liu et al, 2011; Jhuang et al, 2013), demonstrating the benefits of mid-level semantic representations for the analysis of complex human activities. However, all these representations ignore the temporal structure of video, representing actions as orderless feature collections and reducing an *entire* video sequence to an attribute vector. For this reason, we denote them *holistic attribute* representations.

The evolution of semantic concepts has not been thoroughly exploited as a clue for activity understanding, although there have been a few efforts in this direction since our early work of (Li and Vasconcelos, 2012). For example, hidden Markov models (HMM) have been employed to capture the temporal structure of the projection of a video sequence into a space of clusters of visual features (Tang et al, 2012) or a space of supervised attribute detectors (Sun and Nevatia, 2013). Bhattacharya et al (2014) have instead proposed to represent complex activities by the spectrum (or some other harmonic signature) of a model of attribute dynamics derived from the control literature. Finally, Sun and Nevatia (2014) extract discriminative segments from the video and characterize them by temporal transitions of attribute scores.

3 Activity Representation via Attribute Dynamics

In this section, we discuss the representation of activities with attribute dynamics.

3.1 Action Attributes

Attribute representations are members of the class of semantic representations (Rasiwasi et al, 2007; Liu et al, 2011) for image and video. These are representations defined on feature spaces with explicit semantics, *i.e.*, where features are visual concepts, scene classes, *etc.* Images or video are mapped into these spaces by classifiers trained to detect the semantics of interest. For attribute representations, these are binary detectors of video attributes $\{c_k\}_{k=1}^K$ that map a video $\mathsf{v} \in \mathcal{X}$ into a binary vector

$$\boldsymbol{y} = [y_1, \cdots, y_K]^\intercal \in \{0,1\}^K, \qquad (1)$$

indicating the presence/absence of each attribute in v. Classifier output y_k is a Bernoulli random variable, whose probability parameter $\pi_k(\mathsf{v})$ is a confidence score for the presence of attribute c_k in v. This is usually an estimate of the *posterior probability* of attribute c given video v, *i.e.*, $\pi_c(\mathsf{v}) = p(c|\mathsf{v})$. The *semantic space* $\mathcal{S}$ is the space of such scores, defined by

$$\boldsymbol{\pi} : \mathcal{X} \to \mathcal{S} = [0,1]^K, \; \boldsymbol{\pi}(\mathsf{v}) = (\pi_1(\mathsf{v}), \cdots, \pi_K(\mathsf{v}))^\intercal. \qquad (2)$$

The benefits of attribute representations for recognition, namely a higher level of abstraction (which enables better generalization than appearance-based representations), robustness to classification errors, and ability to account for contextual relationships between concepts, have been previously documented in (Lampert et al, 2009; Rasiwasia and Vasconcelos, 2009; Palatucci et al, 2009; Liu et al, 2011; Jhuang et al, 2013).

3.2 Temporal Structure in Attribute Space

Since existing attribute representations do not account for temporal structure, they have limited applicability to video analysis. Temporal structure cannot be captured by representations that are either holistic, such

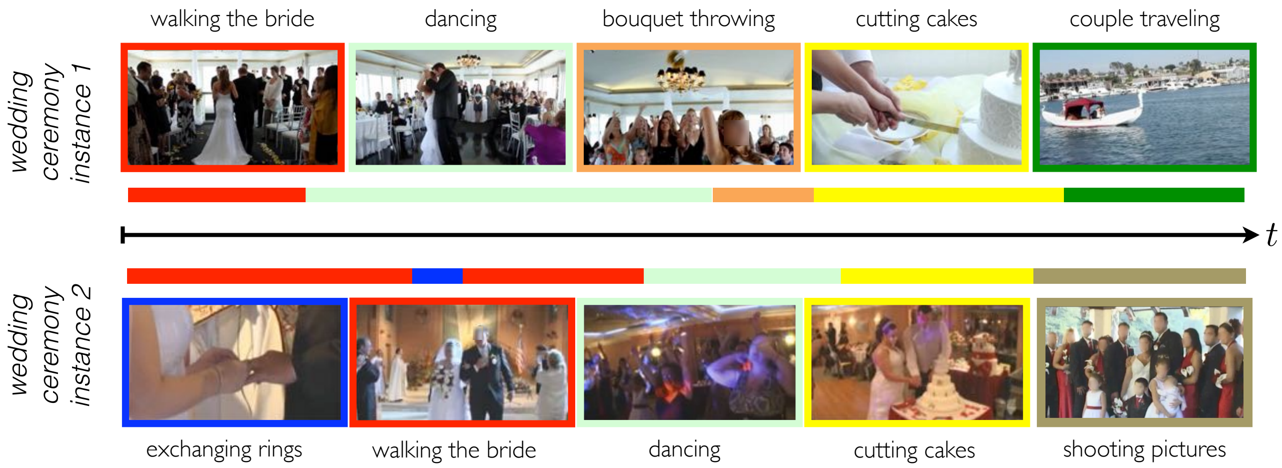


Fig. 3: Video sequences of complex activities, such as "wedding ceremony," are composed by several actions, *e.g.*, "walking the bride", or "cutting cake"). These actions and/or the corresponding durations (indicated by color boxes/bars in the figure) can differ significantly across sequences and are not always informative (*e.g.*, "couple traveling") of the activity class.

as (2), or reduce video to an orderless collection of instantaneous descriptors, such as histograms. We propose to overcome this problem by introducing models of the *dynamics, i.e.*, temporal evolution, of video attributes. This relies on the mapping of each video into a sequence of semantic vectors

$$\boldsymbol{\Pi} = \{\boldsymbol{\pi}_t(\mathsf{v})\} \subset \mathcal{S}, \tag{3}$$

where $\pi_{tk}(\mathsf{v})$ is the confidence score for presence, in v, of attribute k at time t. These scores are obtained by application of attribute detectors to a sliding video window. Fig. 2 motivates the modeling of attribute dynamics, by depicting two activity categories ("long jump" and "hurdle race") that instantiate the same attributes with roughly equal probabilities, but span two very different trajectories in $\mathcal{S}$. While hurdle racing involves a rhythmic transition between short patterns of racing, jumping, and landing, a long jump starts with a longer running sequence, followed by a single jump, and ends with a landing.

It is important to distinguish short- and long-term dynamics. The characterization of short-term dynamics can substantially enhance the expressiveness of a video model. For example, decomposing the activity "long-jump" into the short term events "run-run", "run-jump" and "jump-land", is sufficient to discriminate it from the activity "triple-jump", which is composed of short-term events "run-jump", "jump-jump" and "jump-land". The presence (or absence) of the "jump-jump" segment is the essential difference between the two activities, which are otherwise very similar. In this work, we capture these short-term dynamics with a dynamic Bayesian network, the *binary dynamic system* (BDS), which extends classical linear dynamical systems (Roweis and Ghahramani, 1999) to semantic observations.

Long-term temporal structure, on the other hand, can be less predictable, since attributes of complex activities are highly non-stationary. There are at least three major sources of non-stationarity. First, complex activities are frequently composed of atomic actions with different dynamics. For example, the "wedding ceremony" sequences of Fig. 3 are composed of several events (*e.g.*, "dancing," "cutting the cake," or "bouquet throwing"). Since the dynamics of these events can be quite distinct, it is very challenging to capture the long-term dynamics of the activity with a single model. Second, and more importantly, the training data available is usually too sparse to cover the intra-class variations of high-level activities. For example, while some wedding videos involve scenes of an honeymoon trip, most do not. In this case, attempting to model long-range dynamics is prone to overfitting. Finally, the most discriminant video segments for event recognition are frequently embedded in video that is only marginally informative of the activity class. For example, the discriminant (for weddings) "bouquet toss" sequence can be surrounded by "dancing" sequences (which appear equally in wedding and birthday videos). The ability to identify these discriminant segments, while ignoring the surrounding "action noise" (non-informative segments) are critical for robust event recognition.

These observations suggest that the modeling of dynamics involves a trade-off between gains in discrimination *v.s.* potential for overfitting. Modeling short-term dynamics increases discrimination with small overfitting potential. However, the latter increases with the

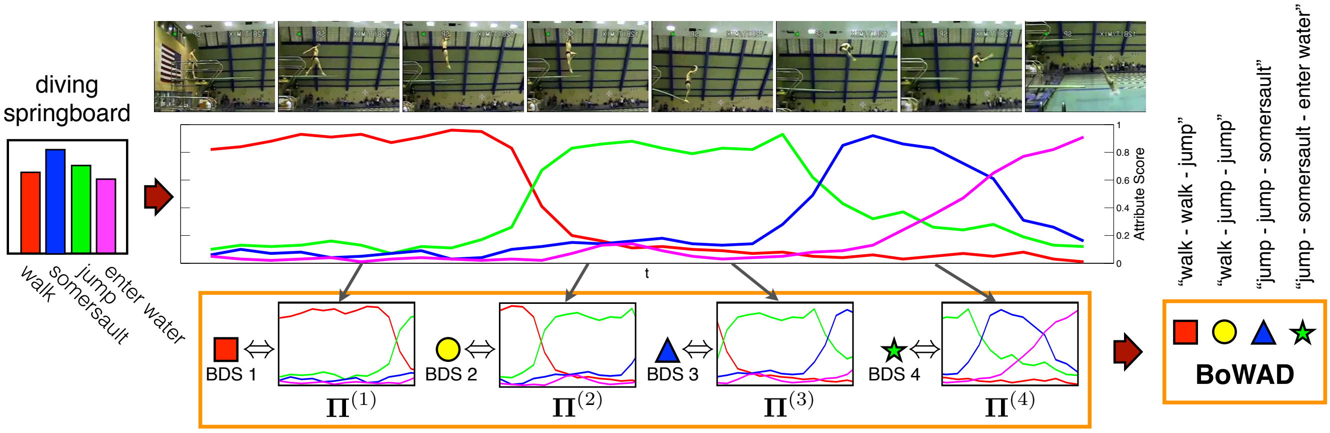


Fig. 4: BoWAD representation of a video of the activity "diving-springboard". (Top) video sequence. (Middle) The classic (holistic) representation of the video on a space of four attributes (represented by four colors) is shown in the left. The proposed representation of the video as a trajectory in the attribute space (four colored functions) is shown at the center. The trajectory is split into overlapping sort-term segments. (Bottom) each segment is assigned to the BDS, in a previously learned dictionary, that best explains it. Dictionary BDS's, denoted WADs, are models of short-term behavior, such as "walk-walk-jump", "walk-jump-jump", "jump-jump-somersault" and "jump-somersault-enter water". The activity is represented by a BoWAD, which is a histogram of assignments of segments to WADs.

temporal support of the video sequences. In result, there is an optimal support, beyond which the benefits of dynamic models start to vanish. This suggests the combination of dynamic models, such as the BDS, for short-term dynamics and representations that may be less discriminant but more robust, such as the BoVW, for long-term dynamics. To accomplish this goal, we propose to encode activity sequences with a BoVW representation that uses the BDS as descriptor of short-term attribute dynamics.

The proposed video representation is illustrated in Fig. 4. A video v is split into segments $\{\boldsymbol{s}^{(i)}\}_{i=1}^N$ of τ_i frames (possibly overlapping in time)[2]. The attribute mapping of (3) is then applied to each segment, producing an attribute sequence $\mathbf{\Pi}^{(i)} = \{\boldsymbol{\pi}_t\}_{t=t_i}^{t_i+\tau_i-1}$, where t_i is the starting time of the i-th segment. v is finally represented by the *bag of attribute sequences* (BoAS) $\{\mathbf{\Pi}^{(i)}\}$ shown in the orange box. This generalizes the BoVW image representation. A dictionary of representative BDSs, denoted *words for attributes dynamics* (WAD), is learned by clustering a collection of BoAS from a set of training attribute sequences. The WAD dictionary is then used to encode the attribute sequences extracted from v as a feature vector for final video classification. This is implemented by either 1) the histogram of WAD counts, denoted a *bag of words for attribute dynamics* (BoWAD), or 2) a descriptor of the first moments of attribute sequences after clustering with a WAD mixture, denoted the *vector of local aggregated descriptors for attribute dynamics* (VLADAD).

4 Models of Attribute Dynamics

In this section, we address the modeling of the dynamics of attribute sequences. We start by considering binary attributes and then generalize the discussion to account for confidence scores.

4.1 Preliminaries

We start by reviewing notation, definitions, and results used throughout this work. We use boldface ($\boldsymbol{x}$) for vectors, capital letters (A) for matrices, $M^\intercal$ to denote the transpose of M, and $\mathrm{tr}(M)$ to denote its trace. $\mathcal{S}^d = \{M|M \in \mathbb{R}^{d\times d}, M = M^\intercal\}$ is the set of $d \times d$ symmetric matrices, and $\mathcal{S}^d_{++} = \{M|M \in \mathcal{S}^d, M \succ \mathbf{0}\}$ the subset of positive-definite matrices.

The probability density (or mass) function of a random vector $\boldsymbol{x}$, with parameter Θ, is denoted $p(\boldsymbol{x};\Theta)$, $p_\Theta(\boldsymbol{x})$, or p_Θ if the argument is clear from context. The

[2] The optimization of the lengths $\{\tau_i\}$ of the video segments $\{\boldsymbol{s}^{(i)}\}$ is left for further research. In this work, we simply considered segments of equal length $\{\tau_i\} = \tau, \forall i$, chosen from a finite set of segment lengths τ, selected so as to achieved good empirical performance on the datasets considered. The specific values of τ used are discussed in the experimental section.

expectation of function $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ is

$$\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x};\Theta)}\left[f(\boldsymbol{x})\right] = \int p(\boldsymbol{x};\Theta)f(\boldsymbol{x})d\boldsymbol{x} = \langle f(\boldsymbol{x})\rangle_{p(\boldsymbol{x};\Theta)} \quad (4)$$

and the Kullback-Leibler (KL) divergence (Kullback, 1997) between distributions $p(\boldsymbol{x};\Theta_1)$ and $p(\boldsymbol{x};\Theta_2)$ is

$$\mathrm{KL}(p_{\Theta_1}||p_{\Theta_2}) = \langle \ln p_{\Theta_1}(\boldsymbol{x})\rangle_{p_{\Theta_1}} - \langle \ln p_{\Theta_2}(\boldsymbol{x})\rangle_{p_{\Theta_1}}. \quad (5)$$

$\mathcal{N}(\boldsymbol{\mu},\Sigma)$ denotes a Gaussian (or normal) distribution with probability density function

$$\mathcal{G}(\boldsymbol{x};\boldsymbol{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\{-\frac{1}{2}||\boldsymbol{x}-\boldsymbol{\mu}||^2_{\Sigma}\}, \quad (6)$$

where d is the dimension of $\boldsymbol{x}$, $\boldsymbol{\mu}\in\mathbb{R}^d$ and $\Sigma\in\mathcal{S}^d_{++}$ are the mean and the covariance, respectively, and

$$||\boldsymbol{x}-\boldsymbol{\mu}||^2_{\Sigma} = (\boldsymbol{x}-\boldsymbol{\mu})^{\intercal}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \quad (7)$$

is the Mahalanobis distance, between $\boldsymbol{x}$ and $\boldsymbol{\mu}$, defined by Σ.

It can be shown (Kullback, 1997) that, when $p_{\Theta_1} = \mathcal{G}(\boldsymbol{x};\boldsymbol{\mu}_1,\Sigma_1)$ and $p_{\Theta_2} = \mathcal{G}(\boldsymbol{x};\boldsymbol{\mu}_2,\Sigma_2)$,

$$\langle \ln p_{\Theta_2}(\boldsymbol{x})\rangle_{p_{\Theta_1}} = \quad (8)$$
$$-\frac{1}{2}\Big[||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2_{\Sigma_2} + d\ln 2\pi + \ln|\Sigma_2| + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1)\Big],$$

and

$$\mathrm{KL}(p_{\Theta_1}||p_{\Theta_2}) = \quad (9)$$
$$\frac{1}{2}\Big[\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + ||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2_{\Sigma_2} - \ln\left|\Sigma_2^{-1}\Sigma_1\right| - d\Big].$$

4.2 Linear Dynamic Systems

Video sequences are frequently modeled as samples of a *linear dynamic system* (LDS)

$$\begin{cases} \boldsymbol{x}_{t+1} = A\boldsymbol{x}_t + \boldsymbol{v}_t, & (10a) \\ \boldsymbol{y}_t = C\boldsymbol{x}_t + \boldsymbol{w}_t + \boldsymbol{u}, & (10b) \end{cases}$$

where $\boldsymbol{x}_t\in\mathbb{R}^L$ and $\boldsymbol{y}_t\in\mathbb{R}^K$ (of mean $\boldsymbol{u}$) are a hidden *state* and *observation* variable at time t, respectively; $A\in\mathbb{R}^{L\times L}$ a state transition matrix that encodes dynamics; $C\in\mathbb{R}^{K\times L}$ an observation matrix that maps state to observations; and $\boldsymbol{x}_1 = \boldsymbol{\mu}+\boldsymbol{v}_0$ an initial condition. Both states and observations have additive Gaussian noise $\boldsymbol{v}_0\sim\mathcal{N}(\boldsymbol{0},S)$, $\boldsymbol{v}_t\sim\mathcal{N}(\boldsymbol{0},Q)$ and $\boldsymbol{w}_t\sim\mathcal{N}(\boldsymbol{0},R)$ $(t\geqslant 1, t\in\mathbb{Z})$. The graphical model of the LDS is shown in Fig. 5.

LDS parameters can be learned by maximum likelihood (ML), using the expectation-maximization (EM) algorithm (Shumway and Stoffer, 1982). A simpler approximate learning procedure was, however, introduced by (Doretto et al, 2003). This is known as the dynamic texture (DT) and decouples the learning of observation and state variables by interpreting the LDS as the combination of a principal component analysis (PCA) and a Gauss-Markov process. Under this interpretation, the columns of C are principal components of the observed video data and the hidden state $\boldsymbol{x}$ is a vector of PCA coefficients. The observation parameters are first learned through a PCA of the video frames, and the state parameters are then learned by least squares. This simple approximate learning algorithm tends to perform very well, and is popular in computer vision.

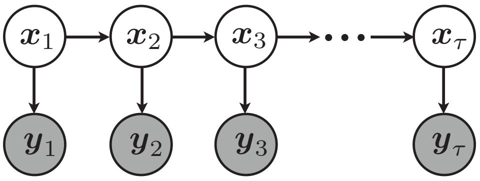


Fig. 5: Graphical model of the LDS and BDS.

4.3 Binary Dynamic Systems

The LDS is a suitable model for the dynamics of continuous observations, such as features extracted from video frames or continuous attributes. However, this is not the common scenario in the attribute literature, where attributes are frequently binary and continuous attributes are usually scores (numbers between 1 and 100), *i.e.*, isomorphic to probabilities. Such attributes clearly violate the assumption of Gaussian observations that underlies the LDS. To address this more challenging (and practically relevant) scenario, we start by introducing an extension of the LDS to sequences $\{\boldsymbol{y}_t\}$ of binary observation vectors. This is denoted the *binary dynamic system* (BDS). The extension to attribute scores is considered in Section 4.5.

The BDS is defined as

$$\begin{cases} \boldsymbol{x}_{t+1} = A\boldsymbol{x}_t + \boldsymbol{v}_t, & (11a) \\ \boldsymbol{y}_t \sim \mathrm{Bern}(\sigma(C\boldsymbol{x}_t+\boldsymbol{u})), & (11b) \end{cases}$$

where $\sigma(\boldsymbol{\theta}) \equiv [\sigma(\theta_1),\cdots,\sigma(\theta_K)]^{\intercal}$, $\sigma(\theta) = (1+e^{-\theta})^{-1}$ is the sigmoid non-linearity, $\mathrm{Bern}(\boldsymbol{\pi})$ the multivariate Bernoulli distribution such that $\boldsymbol{y}\sim\mathrm{Bern}(\boldsymbol{\pi})$,

$$p(\boldsymbol{y};\boldsymbol{\pi}) = \prod\nolimits_k \pi_k^{y_k}(1-\pi_k)^{(1-y_k)}, \quad (12)$$

$\boldsymbol{x}_t\in\mathbb{R}^L$ and $\boldsymbol{u}\in\mathbb{R}^K$ the hidden state variable and observation bias, respectively; $A\in\mathbb{R}^{L\times L}$ a state transition matrix; and $C\in\mathbb{R}^{K\times L}$ an observation matrix. The initial condition is given by $\boldsymbol{x}_1 = \boldsymbol{\mu}+\boldsymbol{v}_0\sim\mathcal{N}(\boldsymbol{\mu},S)$; and the state noise process by $\boldsymbol{v}_t\sim\mathcal{N}(\boldsymbol{0},Q)$. Since the BDS only differs from the LDS in the form of the conditional distribution $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$, its graphical model is identical to that of the LDS, as noted in Fig. 5.

The BDS can be interpreted as a combination of PCA and a Gauss-Markov process by noting that a Bernoulli distribution of parameter π is a member of the exponential family of distributions. Hence, it can be expressed in terms of the natural parameter $\log \frac{\pi}{1-\pi}$, which is mapped to the standard parameter space by the sigmoid $\sigma(\cdot)$. It follows that the BDS represents the video data as an LDS-like observation sequence, $C\boldsymbol{x}_t + \boldsymbol{u}$, in the natural parameter space. Since this is similar to the definition of binary PCA (Schein et al, 2003), the BDS of (11) can be interpreted as the combination of a binary PCA observation component in (11b) and the Gauss-Markov process of (11a). The state vector $\{\boldsymbol{x}_t\}$ thus encodes the trajectory of the binary PCA coefficients of the observed data over time. As is the case for the LDS, this enables an efficient learning algorithm, which we discuss in following sections.

4.4 Binary Principal Component Analysis

Binary PCA (Schein et al, 2003) is a dimensionality reduction technique for binary data, which belongs to the generalized exponential family PCA (Collins et al, 2002). It fits a linear model to binary observations, by embedding the natural parameters of Bernoulli distributions in a low-dimensional subspace. Let Y denote a $K \times \tau$ binary matrix ($y_{kt} \in \{0, 1\}$, *e.g.*, the indicator of occurrence of attribute k at time t) where each column is a vector of K binary observations sampled from a multivariate Bernoulli distribution $Y_{kt} \sim \text{Bern}(\pi_{kt})$ such that

$$p(y_{kt}; \pi_{kt}) = \pi_{kt}^{y_{kt}}(1-\pi_{kt})^{1-y_{kt}} = \sigma(\theta_{kt})^{y_{kt}}\sigma(-\theta_{kt})^{1-y_{kt}} \tag{13}$$

of natural parameters $\theta_{kt} = \log(\frac{\pi_{kt}}{1-\pi_{kt}})$. Binary PCA finds a L-dimensional ($L \ll K$) embedding of the natural parameters, by maximizing the log-likelihood of the binary matrix Y

$$\begin{aligned}\mathcal{L} &= \ln p(\{y_{kt}\}; \Theta) \\ &= \sum_{k,t}\Big[y_{kt} \ln \sigma(\Theta_{kt}) + (1 - y_{kt}) \ln \sigma(-\Theta_{kt})\Big]\end{aligned} \tag{14}$$

under the constraint

$$\Theta = CX + \boldsymbol{u}\boldsymbol{1}^{\intercal}, \tag{15}$$

where $C \in \mathbb{R}^{K\times L}$, $X \in \mathbb{R}^{L\times\tau}$, $\boldsymbol{u} \in \mathbb{R}^K$ and $\boldsymbol{1} \in \mathbb{R}^\tau$ is the vector of all ones. Each column of C is a basis vector of a latent subspace and the t-th column of X contains the coordinates of the t-th binary vector in this basis (up to a translation by $\boldsymbol{u}$).

4.5 Soft Binary PCA

By mapping each video into a sequence of vectors $\{\boldsymbol{\pi}_t\}$ of attribute probabilities, the semantic representation of (3) is much richer than a sequence of binary attribute vectors $\boldsymbol{y}_t$. This, however, prevents the direct application of binary PCA. A solution is nevertheless possible if, instead of the conventional ML criterion, we resort to the maximization of the *expected* log-likelihood of the binary observations $\boldsymbol{y}_t$. This equates parameter learning to the optimization problem

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} \ \langle \ln \mathcal{L}(\boldsymbol{\theta})\rangle_{p(\boldsymbol{y};\boldsymbol{\pi})} &(16)\\ &= \arg\max_{\boldsymbol{\theta}} \ \langle \ln p(Y;\boldsymbol{\theta})\rangle_{p(\boldsymbol{y};\boldsymbol{\pi})}\,. &(17)\end{aligned}$$

Since $\langle \boldsymbol{y}_t\rangle_{p(\boldsymbol{y};\boldsymbol{\pi})} = \boldsymbol{\pi}_t$, it follows from (14) that

$$\langle \mathcal{L}\rangle_{p(\boldsymbol{y};\boldsymbol{\pi})} = \sum_{k,t}\Big[\pi_{kt} \ln \sigma(\Theta_{kt}) + (1 - \pi_{kt}) \ln \sigma(-\Theta_{kt})\Big], \tag{18}$$

and (17) can be solved with the binary PCA algorithm.

It should be noted that this solution is identical to the ML estimate of binary PCA in the case of infinite data since, by the law of large numbers,

$$\frac{1}{N}\sum_{i=1}^{N} \ln p(y^{(i)}; \boldsymbol{\theta}) \xrightarrow[N\to\infty]{} \langle \ln p(Y;\boldsymbol{\theta})\rangle_{p(\boldsymbol{y};\boldsymbol{\pi})}\,,$$

where $\{y^{(i)}\}_{i=1}^N$ are N *independent and identically distributed* (*i.i.d.*) examples from $p(\boldsymbol{y}; \boldsymbol{\pi})$. The solution of (17) also minimizes the KL divergence between $p(\boldsymbol{y}; \boldsymbol{\pi})$ and the model $p(\boldsymbol{y}; \boldsymbol{\theta})$, since

$$\begin{aligned}&\mathrm{KL}(p(\boldsymbol{y}; \boldsymbol{\pi})||p(\boldsymbol{y}; \boldsymbol{\theta})) &(19)\\ &\quad= \ \langle \ln p(Y; \boldsymbol{\pi})\rangle_{p(\boldsymbol{y};\boldsymbol{\pi})} - \langle \ln p(Y; \boldsymbol{\theta})\rangle_{p(\boldsymbol{y};\boldsymbol{\pi})} \geqslant 0,\end{aligned}$$

and the first term is independent of $\boldsymbol{\theta}$.

4.6 BDS Learning

The discussion above suggests a generalization of the DT learning procedure to the BDS. The soft binary PCA basis is learned first, by maximizing the expected log-likelihood of (18) subject to the constraint of (15). Since the Bernoulli is a member of exponential family, (18) is concave in Θ, but not in C, X and $\boldsymbol{u}$ jointly. The ML parameters can be found with the procedure of (Schein et al, 2003), which iterates between the optimization with respect to one of the variables C, X and $\boldsymbol{u}$ as the other two are held constant. Each iteration is a convex sub-problem that can be solved efficiently with a fixed-point auxiliary function (Schein et al, 2003).

Algorithm 1: BDS learning

```
Input  : a set of $n$ sequences of attribute score vectors
         $\{\{\boldsymbol{\pi}_t^{(i)}\}_{t=1}^{\tau_i}\}_{i=1}^{n}$, state space dimension $L$.
  Soft binary PCA (Schein et al, 2003):
  $\{C, X, \boldsymbol{u}\} = \text{B-PCA}(\{\{\boldsymbol{\pi}_t^{(i)}\}_{t=1}^{\tau_i}\}_{i=1}^{n}, L)$;
  Assemble state sequences $(X_{t_1}^{t_2} \equiv [\boldsymbol{x}_{t_1}, \cdots, \boldsymbol{x}_{t_2}])$ :
     $\hat{X}_2^{\tau} = [(X^{(1)})_2^{\tau_1}, \cdots, (X^{(n)})_2^{\tau_n}]$,
  $\hat{X}_1^{\tau-1} = [(X^{(1)})_1^{\tau_1 - 1}, \cdots, (X^{(n)})_1^{\tau_n - 1}]$;
  Estimate state parameters:
  $A = \hat{X}_2^{\tau}(\hat{X}_1^{\tau-1})^{\dagger}, \quad V = \hat{X}_2^{\tau} - A\hat{X}_1^{\tau-1}$,
  $Q = \frac{1}{\sum_i(\tau_i - 1)} V(V)^{\intercal}, \quad \boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_1^{(i)}$,
  $S = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_1^{(i)} - \boldsymbol{\mu})(\boldsymbol{x}_1^{(i)} - \boldsymbol{\mu})^{\intercal}$.
Output: $\boldsymbol{\Omega} = \{A, C, Q, \boldsymbol{u}, \boldsymbol{\mu}, S\}$
```

Once the optimal embedding C^*, X^* and $\boldsymbol{u}^*$ of the attribute sequence is recovered, the remaining parameters are estimated by solving a least-squares problem for A and Q, and using ML estimates for the Gaussian parameters of the initial condition ($\boldsymbol{\mu}_0$ and S_0). Since this is identical to the least squares procedure of (Doretto et al, 2003), we omit the details. The learning procedure, including the least squares equations, is summarized in Algorithm 1. Since the optimal solution maximizes the most natural measure of similarity (KL divergence) between probability distributions, this extension is conceptually equivalent to the procedure used to learn the LDS, which finds the subspace that best fits the observations in the Euclidean sense, the natural similarity measure for Gaussian data. This is unlike previous extensions of the LDS, *e.g.*, kernel dynamic systems (KDS) that rely on a non-linear kernel PCA (KPCA) (Schölkopf et al, 1998) of the observation space but still assume an Euclidean measure (Gaussian noise) (Chan and Vasconcelos, 2007; Chaudhry et al, 2009). In the experimental section we show that the BDS is a superior model of attribute dynamics.

5 Bag-of-Words for Attribute Dynamics

In this section, we introduce the *bag-of-words for attribute dynamics* (BoWAD) representation.

5.1 Clustering Samples in the Model Domain

Clustering identify prototypes in the space of training examples (*e.g.*, in k-means, a cluster prototype is the centroid of the samples in the cluster), using a metric suited for that space (*e.g.*, Euclidean distance). Clustering BoAS is not straightforward because 1) attribute sequences can have different length; 2) the space of these sequences has non-Euclidean geometry; and 3) the search for optimal prototypes, under this geometry, may lead to intractable non-linear optimization. This is compounded by the fact that the dynamics of attribute sequences are better summarized by a set of prototype BDSs than a set of prototype sequences.

Algorithm 2: Bag-of-Models Clustering

```
Input  : a set of samples $\mathcal{D} = \{\boldsymbol{z}_i\}_{i=1}^{N}$ $(\boldsymbol{z}_i \in \mathcal{Z}, \forall i)$,
         number of clusters $N_C$, an initial set of
         models $\{M_i^{(0)}\}_{i=1}^{N_C}$.
set $t = 0$ and $S_i^{(0)} = \varnothing, i = 1, \cdots, N_C$;
repeat
    $t = t + 1$;
    Assignment-Step:  $\forall i,\ S_i^{(t)} = \{\boldsymbol{z} \in \mathcal{D} \mid \forall j \neq i,$
     $d_{\mathcal{M}}(M(\boldsymbol{z}), M_i^{(t-1)}) \leqslant d_{\mathcal{M}}(M(\boldsymbol{z}), M_j^{(t-1)})\}$;
    Refinement-Step:  $\forall i,\ M_i^{(t)} = M(S_i^{(t)})$;
until $\forall i, S_i^{(t)} = S_i^{(t-1)}$;
Output: $\{M_i^{(t)}\}_{i=1}^{N_C}$ and $\{S_i^{(t)}\}_{i=1}^{N_C}$
```

The problem of learning a set of BDS prototypes is an instance of the problem of learning a *bag-of-models* (BoM). Given a training set $\mathcal{D} = \{\boldsymbol{z}_i\}_{i=1}^{N}$ $(\boldsymbol{z}_i \in \mathcal{Z}, \forall i)$, the goal is to learn a dictionary of representative *models* $\{M_i(\boldsymbol{z})\}_{i=1}^{N_C}$ in a *Riemannian manifold* $\mathcal{M}$ of models. The proposed solution is based on two mappings. The first

$$f_{\mathcal{M}} : \mathcal{Z} \supseteq \{\boldsymbol{z}_i\} \mapsto M \in \mathcal{M} \tag{20}$$

maps a set of examples $\{\boldsymbol{z}_i\} \subseteq \mathcal{D}$ into a model $M(\boldsymbol{z})$. The second,

$$\mathcal{M} \times \mathcal{M} \ni (M_1, M_2) \mapsto d_{\mathcal{M}}(M_1, M_2) \in \mathbb{R}_+ \tag{21}$$

measures the dissimilarity or distance between models (*e.g.*, geodesic distance in the manifold).

The mapping of (20) is first used to produce a model $M(\boldsymbol{z}_i)$ per training example $\boldsymbol{z}_i$. Training samples are then clustered, at the model level, by alternating between two steps. In the *assignment step*, each $\boldsymbol{z}_i$ is assigned to the cluster whose model is closest to $M(\boldsymbol{z}_i)$, using the mapping of (21). In the *model refinement step*, the model associated with each cluster is relearned from the training samples assigned to it, via (20). This procedure is summarized in Algorithm 2 and denoted *bag-of-models clustering* (BMC). It can be shown that, under some mild conditions, it converges in a finite number steps. A sketch of the proof is provided in Appendix A.

BMC generalizes k-means, where $\boldsymbol{z}_i \in \mathbb{R}^d$ are feature vectors, $\mathcal{M}$ is the space of Gaussians of identity covariance

$$\mathcal{M} = \left\{\mathcal{G}(\boldsymbol{z}; \boldsymbol{\mu}, I_d) \mid \boldsymbol{\mu} \in \mathbb{R}^d \right\}, \tag{22}$$

(20) selects the model

$$M(\{\boldsymbol{z}_i\}) = \mathcal{G}(\boldsymbol{z}; \hat{\boldsymbol{\mu}}, I), \tag{23}$$

where $\hat{\boldsymbol{\mu}}$ is the ML estimate of the mean

$$\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} p(\{\boldsymbol{z}_i\}; \boldsymbol{\mu}) = \frac{1}{|\{\boldsymbol{z}_i\}|} \sum\nolimits_i \boldsymbol{z}_i, \tag{24}$$

and (21) is the symmetric KL divergence derived from (9),

$$\mathrm{KL}(p_1||p_2) + \mathrm{KL}(p_2||p_1) = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2. \tag{25}$$

It should be noted that BMC differs from the *bag-of-systems* approach (Ravichandran et al, 2012; Afsari et al, 2012) in two ways. First, it clusters *attribute sequences* rather than models. While, in the refinement step of Algorithm 2, models are re-learned from examples $\{\boldsymbol{z}_i\}$, the refinement step of (Ravichandran et al, 2012; Afsari et al, 2012) only considers parameters of the models $M(\boldsymbol{z}_i)$ and not the examples $\boldsymbol{z}_i$ themselves. This usually entails loss of information. Second, Algorithm 2 *finds* the optimal representative for each cluster, according to the model fitting criterion of (20). In (Ravichandran et al, 2012), the difficult geometry of the manifold defined by the LDS parameter tuple $(A, C) \in \mathbb{GL}(n) \times \mathbb{ST}(p, n)$, where $\mathbb{GL}(i)$ is the set of invertible matrices of size n and $\mathbb{ST}(p, n)$ the Stiefel manifold of $p \times n$ orthonormal matrices ($p \geqslant n$), precludes a simple estimate of the optimal representative. Instead, this is approximated by model $M(\boldsymbol{z}_i)$ closest to the optimal representative. Although Afsari et al (2012) introduce an approach to directly cluster LDS's in parameter space, its generalization to the BDS is unclear. We will show, in Section 7, that these differences can lead to significantly improved performance by Algorithm 2.

5.2 Dissimilarity Measure Between BDSs

Algorithm 2 requires a measure of distance Between BDSs. For this, we generalize a popular measure of distance between LDSs, the Binet-Cauchy kernel (BCK) of (Vishwanathan et al, 2006). Given LDSs $\boldsymbol{\Omega}_a$ and $\boldsymbol{\Omega}_b$ driven by identical noise processes $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ with observation sequences $\boldsymbol{y}^{(a)}$ and $\boldsymbol{y}^{(b)}$, the BCK is

$$K_{BC}(\boldsymbol{\Omega}_a, \boldsymbol{\Omega}_b) = \left\langle \sum_{t=0}^{\infty} e^{-\lambda t} (\boldsymbol{y}_t^{(a)})^\intercal W \boldsymbol{y}_t^{(b)} \right\rangle_{p(\boldsymbol{v}, \boldsymbol{w})}, \tag{26}$$

where W is a semi-definite positive weight matrix and $\lambda \geqslant 0$ a temporal discounting factor. To extend (26) to BDSs $\boldsymbol{\Omega}_a$ and $\boldsymbol{\Omega}_b$, we note that $(\boldsymbol{y}_t^{(a)})^\intercal W \boldsymbol{y}_t^{(b)}$ is the inner product of the Euclidean space of metric $d^2(\boldsymbol{y}_t^{(a)}, \boldsymbol{y}_t^{(b)}) = (\boldsymbol{y}_t^{(a)} - \boldsymbol{y}_t^{(b)})^\intercal W (\boldsymbol{y}_t^{(a)} - \boldsymbol{y}_t^{(b)})$. For BDSs, whose observations $\boldsymbol{y}_t$ are Bernoulli distributed with parameters $\{\sigma(\boldsymbol{\theta}_t^{(a)})\}$, for $\boldsymbol{\Omega}_a$, and $\{\sigma(\boldsymbol{\theta}_t^{(b)})\}$, for $\boldsymbol{\Omega}_b$, this distance measure is naturally replaced by the symmetric KL divergence between Bernoulli distributions. This results in the *Binet-Cauchy KL divergence* (BC-KLD)

$$\begin{aligned} &D_{BC}(\boldsymbol{\Omega}_a, \boldsymbol{\Omega}_b) \\ &\quad = \left\langle \sum_{t=0}^{\infty} e^{-\lambda t} \Big[\mathrm{KL}(B(\sigma(\boldsymbol{\theta}_t^{(a)}))||B(\sigma(\boldsymbol{\theta}_t^{(b)}))) \right. \\ &\qquad\qquad \left. + \mathrm{KL}(B(\sigma(\boldsymbol{\theta}_t^{(b)}))||B(\sigma(\boldsymbol{\theta}_t^{(a)}))) \Big] \right\rangle_{p(\boldsymbol{v})} \\ &\quad = \left\langle \sum_{t=0}^{\infty} e^{-\lambda t} \left[\sigma(\boldsymbol{\theta}_t^{(a)}) - \sigma(\boldsymbol{\theta}_t^{(b)}) \right]^\intercal \left[\boldsymbol{\theta}_t^{(a)} - \boldsymbol{\theta}_t^{(b)} \right] \right\rangle_{p(\boldsymbol{v})}, \end{aligned} \tag{27}$$

where $\boldsymbol{\theta}_t = C\boldsymbol{x}_t + \boldsymbol{u}$ is the parameter of the multivariate Bernoulli distribution. [3] The divergence at time t can be rewritten as

$$\begin{aligned} &(\sigma(\boldsymbol{\theta}_t^{(a)}) - \sigma(\boldsymbol{\theta}_t^{(b)}))^\intercal (\boldsymbol{\theta}_t^{(a)} - \boldsymbol{\theta}_t^{(b)}) \\ &\quad = (\boldsymbol{\theta}_t^{(a)} - \boldsymbol{\theta}_t^{(b)})^\intercal \hat{W}_t (\boldsymbol{\theta}_t^{(a)} - \boldsymbol{\theta}_t^{(b)}), \end{aligned} \tag{28}$$

with $\hat{W}_t$ a diagonal matrix whose k-th diagonal element is $\hat{W}_{t,k} = (\sigma(\theta_{t,k}^{(a)}) - \sigma(\theta_{t,k}^{(b)}))/(\theta_{t,k}^{(a)} - \theta_{t,k}^{(b)}) = \sigma'(\hat{\theta}_{t,k}^{(a,b)})$ (where, by the mean value theorem, $\hat{\theta}_{t,k}^{(a,b)}$ is some real value between $\hat{\theta}_{t,k}^{(a)}$ and $\hat{\theta}_{t,k}^{(b)}$). This reduces (28) to a form similar to (26), although with a time varying weight matrix W_t. It is, nevertheless unclear whether (27) can be computed in closed-form. We rely on the approximation

$$\begin{aligned} &D_{BC}(\boldsymbol{\Omega}_a, \boldsymbol{\Omega}_b) \\ &\approx \sum_{t=0}^{\infty} e^{-\lambda t} \left[\sigma(\bar{\boldsymbol{\theta}}_t^{(a)}) - \sigma(\bar{\boldsymbol{\theta}}_t^{(b)}) \right]^\intercal \left[\bar{\boldsymbol{\theta}}_t^{(a)} - \bar{\boldsymbol{\theta}}_t^{(b)} \right], \end{aligned} \tag{29}$$

where $\bar{\boldsymbol{\theta}}$ is the mean of $\boldsymbol{\theta}$.

5.3 Learning a WAD Vocabulary

Given the BC-KLD distance between BDSs, it is possible to learn a WAD dictionary from a BoAS $\mathcal{P} = \{\boldsymbol{\Pi}^{(i)}\}_{i=1}^N$, by applying Algorithm 2 as follows.

Refinement-Step: The mapping of (20) amounts to fitting a BDS to a BoAS $\mathcal{P}' = \{\boldsymbol{\Pi}^{(i)}\} \subseteq \mathcal{P}$. This is

[3] Although the square root of the symmetric KL divergence is not a metric (since the triangle inequality does not hold), it has been shown effective for the design of probability distribution kernels, in the context of various applications (Moreno et al, 2004; Vasconcelos et al, 2004; Haasdonk, 2005; Chan and Vasconcelos, 2005).

done with Algorithm 1. The BDS learned per cluster jointly characterizes the appearance and dynamics of all attribute sequences in that cluster.

Assignment-Step: Each sample BDS is assigned to the closest centroid BDS, using (29).

To initialize the clustering algorithm, we follow the strategy of (Chan and Vasconcelos, 2008). This has produced satisfactory results in all our experiments.

5.4 Quantization of BoAS with WAD Vocabulary

Given a WAD dictionary $\{\boldsymbol{\Omega}^{(i)}\}_{i=1}^{V}$, a BoAS $\{\{\boldsymbol{\pi}_t\}_{t=t_i}^{t_i+\tau_i-1}\}$ is quantized by assigning the i-th attribute sequence to the k^*-th cluster according to

$$k^* = \arg\min_j \, d_{BC}\big(\boldsymbol{\Omega}(\{\boldsymbol{\pi}_t\}_{t=t_i}^{t_i+\tau_i-1}), \boldsymbol{\Omega}^{(j)}\big), \tag{30}$$

where $\boldsymbol{\Omega}(\{\boldsymbol{\pi}_t\}_{t=t_i}^{t_i+\tau_i-1})$ is the BDS learnt from $\{\boldsymbol{\pi}_t\}_{t=t_i}^{t_i+\tau_i-1}$ using (20). This produces a histogram of WAD counts, denoted *bag-of-words for attribute dynamics* (BoWAD), which can be used to classify video sequences of complex activities with the procedures commonly used for the BoVW (Laptev et al, 2008; Wang et al, 2009).

6 The VLAD for Attribute Dynamics

In this section, we derive a VLAD encoding for attribute dynamics.

6.1 Bag-of-Models Interpretation of VLAD

The VLAD is an efficient representation of the first moments of a data sample. It has been shown to outperform the BoVW histogram, which only captures zeroth moments, in many image classification experiments. To extend the VLAD to the BoM, we start by interpreting it as an encoding of sample statistics with respect to a collection of local tensors of a model manifold.

Consider a Riemannian manifold $\mathcal{M}$ with geodesic distance $d_{\mathcal{M}}(M_1, M_2)$, such as (21), a set of reference models $\{M_i\}_{i=1}^{N_C}$, embedded in $\mathcal{M}$, and neighborhoods

$$\mathcal{R}_i = \{M \in \mathcal{M} | d_{\mathcal{M}}(M, M_i) \leqslant d_{\mathcal{M}}(M, M_j), j \neq i\},$$

where $\mathcal{R}_i$ is the neighborhood of M_i under $d_{\mathcal{M}}$. To encode a collection of examples $\mathcal{D} = \{\boldsymbol{z}_i\}_{i=1}^{N}$ ($\boldsymbol{z}_i \in \mathcal{Z}, \forall i$), these are first assigned to the regions $\mathcal{R}_i$

$$\mathcal{D}^i = \{\boldsymbol{z} \in \mathcal{D} | f_{\mathcal{M}}(\boldsymbol{z}) \in \mathcal{R}_i\} \tag{31}$$

using an assignment mapping $f_{\mathcal{M}}$, such as (20).

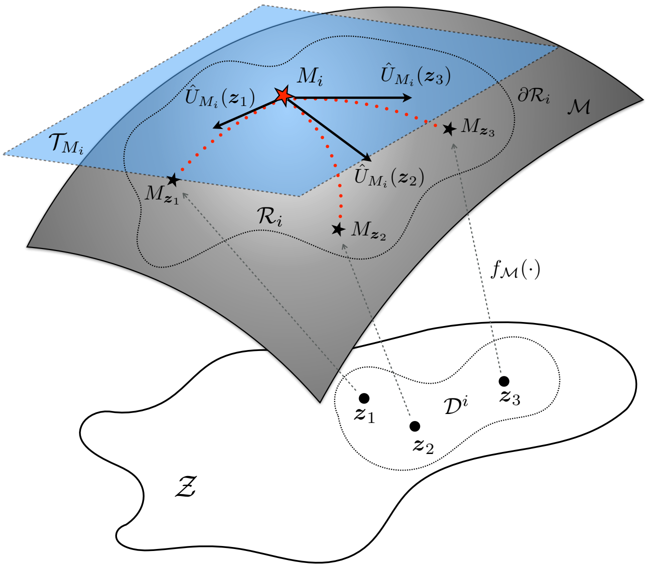


Fig. 6: VLAD encoding under the bag of models representation. The data points in $\mathcal{D}^i$ are first mapped into model manifold $\mathcal{M}$ by $f_{\mathcal{M}_i}(\boldsymbol{z})$, and then encoded by their first moments with respect to M_i (the red star in the figure) to approximate the geodesics (*e.g.*, the geodesic distances in red dotted curves), using the mapping $\hat{U}_{M_i}(\boldsymbol{z}) = \mathcal{I}_{M_i}^{-1/2} U_{M_i}(\boldsymbol{z})$ defined by the local tensor $\mathcal{T}_{M_i}$, *i.e.*, the metric of the tangent space at M_i (the blue plane in the figure).

VLAD assumes examples $\boldsymbol{z} \in \mathbb{R}^D$ and Gaussian models M_i, *i.e.*, a model manifold

$$\mathcal{M} = \big\{\mathcal{G}(\boldsymbol{z}; \boldsymbol{\mu}, \Sigma) \mid \boldsymbol{\mu} \in \mathbb{R}^D, \Sigma \in \mathcal{S}_{++}^D \big\}, \tag{32}$$

with geodesic distance approximated by the symmetric KL divergence

$$d_{\mathcal{M}}(M_1, M_2) = \mathrm{KL}(p_{M_1} || p_{M_2}) + \mathrm{KL}(p_{M_2} || p_{M_1}), \tag{33}$$

where $\mathrm{KL}(p_{M_1} || p_{M_2})$ is defined in (9). Most VLAD implementations assume that $\Sigma = I$, reducing (33) to the Euclidean metric $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ ($\mathcal{D}^i$ assigned to the model of mean closest to the sample centroid). In this case, the assignment mapping maps an example $\boldsymbol{z}$ to a Gaussian of mean $\boldsymbol{\mu}$ and identity covariance, *i.e.*,

$$f_{\mathcal{M}}(\boldsymbol{z}) : \boldsymbol{z} \rightarrow \mathcal{G}(\boldsymbol{z}; \boldsymbol{\mu}, I_D), \tag{34}$$

where $\boldsymbol{\mu} \in \{\boldsymbol{\mu}_i\}$ is the mean of one of the reference Gaussians.

As illustrated in Fig. 6, the idea behind VLAD is to use the local tensor $\mathcal{T}_{M_i}$ defined by distance $d_{\mathcal{M}}(\cdot, \cdot)$ at M_i to encode the distribution of $\mathcal{D}^i$. A descriptor of $\mathcal{D}$ is then constructed by 1) aggregating the encoding of the examples in $\mathcal{D}^i$, for each region $\mathcal{R}_i$, and 2) concatenating the aggregate encodings from all regions. When $\mathcal{M}$

is a statistical manifold (of parameter $\boldsymbol{\theta}$), a commonly used metric tensor is the *Fisher kernel* (Jaakkola and Haussler, 1999)

$$K_M(\boldsymbol{z}_1, \boldsymbol{z}_2) = U_M^{\intercal}(\boldsymbol{z}_1)\mathcal{I}_M^{-1}U_M(\boldsymbol{z}_2), \tag{35}$$

where

$$U_M(\boldsymbol{z}) = \nabla_{\boldsymbol{\theta}} \log p_M(\boldsymbol{z}; \boldsymbol{\theta}) \tag{36}$$

is the *Fisher score* and $\mathcal{I}_M$ is the *Fisher information metric* at M. [4] This tensor can be shown to approximate the KL-divergence in the neighborhood of M (Amari, 1998; Amari and Nagaoka, 2000).

For the manifold of (32), the Fisher score is

$$U_M(\boldsymbol{z}) = \begin{bmatrix} \nabla_{\boldsymbol{\mu}} \log p_M(\boldsymbol{z}; \boldsymbol{\mu}, \Sigma) \\ \nabla_{\Sigma^{-1}} \log p_M(\boldsymbol{z}; \boldsymbol{\mu}, \Sigma) \end{bmatrix},$$

with

$$\nabla_{\boldsymbol{\mu}} \log p_M(\boldsymbol{z}) = \Sigma^{-1}(\boldsymbol{z} - \boldsymbol{\mu}), \tag{37}$$

$$\nabla_{\Sigma^{-1}} \log p_M(\boldsymbol{z}) = \frac{1}{2}\Big[\Sigma - (\boldsymbol{z} - \boldsymbol{\mu})(\boldsymbol{z} - \boldsymbol{\mu})^{\intercal}\Big]. \tag{38}$$

After the aggregation over the sample $\mathcal{D}^i$, (37) encodes the relative position of the centroid of this sample w.r.t. the region center $\boldsymbol{\mu}_i$ (under the Mahalanobis metric defined by Σ_i^{-1}). Similarly, (38) encodes the relative shape of the sample w.r.t. that of the reference distribution, which is parametrized by Σ_i. Under the assumption that $\Sigma = I$, (37) reduces to $\boldsymbol{z} - \boldsymbol{\mu}$ and the second moments of (38) are usually omitted. This has some loss but reduces complexity (Jegou et al, 2012).

6.2 Variational Inference for the BDS

The extension of the VLAD to the BDS requires evaluating the derivative of the expected log-likelihood of the sample with respect to the model parameters. This, however, is intractable, due to the intractability of the posterior distribution of BDS state given observations. To overcome this difficulty, we resort to approximate variational inference (Jordan et al, 1999). A similar strategy has recently been shown effective for image analysis (Cinbis et al, 2012).

6.2.1 Variational Inference

Given a a model $p(Y, X; \theta)$ with parameter θ, observed variable Y and hidden variable X, there is usually a need to evaluate the posterior distribution $p(X|Y; \theta)$. While this is tractable for some dynamic models, *e.g.*, the LDS of (10) where the Gaussian hidden state distribution is a conjugate prior for the Gaussian conditional-distribution of observations given state, it is intractable for the BDS of (11), where the state is Gaussian but the observations are not. Variational inference is a tool for approximate inference in problems with this type of intractability.

Variational methods are based on the decomposition of the marginal likelihood of the observed data

$$\begin{aligned} \ln \mathcal{L}(\theta, y) &= \ln p(y; \theta) && (39)\\ &= \mathscr{L}(\theta, y, q) + \mathrm{KL}(q(x)||p(x|y; \theta)), && (40) \end{aligned}$$

where

$$\mathscr{L}(\theta, y, q) = \int_x q(x) \ln \frac{p(y, x; \theta)}{q(x)} dx \tag{41}$$

and $q(x)$ some probability distribution. Since the KL divergence is non-negative, $\mathscr{L}(\theta, y, q)$ is a lower bound of $\ln \mathcal{L}(\theta; y)$. For a family $\mathcal{D}_q$ of tractable distributions of X, the tightest lower bound is achieved at

$$q_y^*(x) = \arg\max_{q \in \mathcal{D}_q} \mathscr{L}(\theta, y, q). \tag{42}$$

This also minimizes the distance to the posterior $p(x|y; \theta)$, in the KL sense, since $\ln \mathcal{L}(\theta, y)$ does not depend on $q(x)$. Hence, the intractable posterior $p(x|y; \theta)$ can be replaced by the variational distribution $q(x)$, and the tighest bound used as a proxy for the log-likelihood $\ln \mathcal{L}(\theta, y)$, for the purpose of learning the model parameters.

6.2.2 Variational Inference for Expected Log-likelihood

The variational setting for learning BDS parameters is slightly different from the standard variational setting because, in (17), the goal is to maximize the expected log-likelihood with regards to a reference distribution $\tilde{p}(y) = p(\boldsymbol{y}; \boldsymbol{\pi})$, *i.e.*

$$\langle \ln \mathcal{L}(\theta, y) \rangle_{\tilde{p}(y)} = \langle \ln p(y; \theta) \rangle_{\tilde{p}(y)}. \tag{43}$$

In this case

$$\begin{aligned} \langle \ln \mathcal{L}(\theta, y) \rangle_{\tilde{p}(y)} &= \mathscr{L}(\theta, q) + \langle \mathrm{KL}(q(x)||p(x|y; \theta)) \rangle_{\tilde{p}(y)} \\ &\geqslant \mathscr{L}(\theta, q) && (44) \end{aligned}$$

[4] In practice, the Fisher information metric $\mathcal{I}_M$ is often omitted, since the Fisher kernel is an Euclidean metric in the range space of the invertible linear transformation by $\mathcal{I}_M^{1/2}$, of the tangent space of the manifold at M.

with lower bound

$$\mathscr{L}(\theta, q) = \langle \mathscr{L}(\theta, y, q)\rangle_{\tilde{p}(y)} \tag{45}$$

$$= \int_x q(x)\,\langle \ln p(y, x; \theta)\rangle_{\tilde{p}(y)}\, dx + H_q(X), \tag{46}$$

where $H_q(X) = -\int_x q(x)\ln q(x)dx$ is the entropy of X under distribution $q(x)$. This bound is tightest at

$$q^*(x) = \arg\max_{q\in\mathcal{D}_q} \mathscr{L}(\theta, q) \tag{47}$$

$$= \arg\min_{q\in\mathcal{D}_q} \langle \mathrm{KL}(q(x)||p(x|y;\theta))\rangle_{\tilde{p}(y)}\,. \tag{48}$$

Note that, by Jensen's inequality,

$$\mathscr{L}(\theta, q^*) = \max_{q\in\mathcal{D}_q} \langle \mathscr{L}(\theta, y, q)\rangle_{\tilde{p}(y)} \tag{49}$$

$$\leqslant \left\langle \max_{q\in\mathcal{D}_q} \mathscr{L}(\theta, y, q)\right\rangle_{\tilde{p}(y)} \tag{50}$$

$$= \left\langle \mathscr{L}(\theta, y, q_y^*)\right\rangle_{\tilde{p}(y)}\,. \tag{51}$$

Hence, the tightest bound of the expected log-likelihood lower bounds the average tightest log-likelihood bounds across observation sequences. Intuitively, (49) lower bounds the log-likelihood over all samples from $\tilde{p}(y)$ that share the same hidden variable, distributed according to $q^*(x)$. On the other hand, (51) uses the distribution $q_y^*(x)$ that best explains each sample y.

6.2.3 Variational Distribution for the BDS

Under the BDS, the log-probability of the complete data is

$$\ln p(\boldsymbol{x}_1^\tau, \boldsymbol{y}_1^\tau; \boldsymbol{\theta})$$
$$= \ln p(\boldsymbol{x}_1) + \sum_{t=1}^{\tau-1} \ln p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) + \sum_{t=1}^{\tau} \ln p(\boldsymbol{y}_t|\boldsymbol{x}_t), \tag{52}$$

where

$$p(\boldsymbol{x}_1) = \mathcal{G}(\boldsymbol{x}_1; \boldsymbol{\mu}, S), \tag{53}$$

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \mathcal{G}(\boldsymbol{x}_{t+1}; A\boldsymbol{x}_t, Q), \tag{54}$$

$$\ln p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \sum_{k=1}^{K} \Big[\, y_{kt}\ln\sigma(\omega_{kt}) + (1-y_{kt})\ln\sigma(-\omega_{kt})\Big], \tag{55}$$

$$\omega_{kt} = C_{k\cdot}\boldsymbol{x}_t + u_k, \tag{56}$$

and $C_{k\cdot}$ is the k-th row of C in (11). The mixture of quadratic and log-sigmoid terms in (52) makes the evaluation of $p(\boldsymbol{x}_1^\tau|\boldsymbol{y}_1^\tau;\boldsymbol{\theta})$ intractable. A family $\mathcal{D}_q$ of tractable variational distributions is required to derive the variational lower bound $\mathscr{L}(\theta, q^*)$ of (49). The most popular strategy in the literature is to adopt factorized variational distributions $q(\boldsymbol{x}) = \prod_{t=1}^{\tau} q_t(\boldsymbol{x}_t)$ (Attias, 1999; Ghahramani and Beal, 2000; Winn and Bishop, 2005) for computational tractability. This, however, fails to capture the dependency among hidden variables, defeating the purpose of dynamic modeling by the BDS. To avoid this problem, we adopt a multivariate Gaussian distribution of *full* covariance for $q(\boldsymbol{x})$,

$$q(\boldsymbol{x}_1^\tau) = \mathcal{G}(\boldsymbol{x}_1^\tau; \boldsymbol{m}, \Sigma),\;\; \boldsymbol{m}\in\mathbb{R}^{L\tau\times 1},\; \Sigma\in\mathcal{S}_{++}^{L\tau}, \tag{57}$$

where $\boldsymbol{m}_i \in \mathbb{R}^L$ and $\Sigma_{i,j}\in\mathbb{R}^{L\times L}$ are the mean of $\boldsymbol{x}_i$ and covariance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively,

$$\boldsymbol{m}_i = \langle \boldsymbol{x}_i\rangle_q\,,\quad \Sigma_{i,j} = \langle(\boldsymbol{x}_i - \boldsymbol{m}_i)(\boldsymbol{x}_j - \boldsymbol{m}_j)^\intercal\rangle_q\,.$$

Given a reference distribution for the sequence of binary attribute indicators $\boldsymbol{y}_1^\tau$, $\tilde{p}(\boldsymbol{y}) = p(\boldsymbol{y};\boldsymbol{\pi}) = p(\boldsymbol{y};\{\pi_{k,t}\})$, it follows from (46) and (52) that

$$\mathscr{L}(\boldsymbol{\theta}, q) = \langle \ln p(\boldsymbol{x}_1)\rangle_q + \sum_{t=1}^{\tau-1}\langle \ln p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)\rangle_q$$
$$+ \sum_{t=1}^{\tau}\Big\langle \langle \ln p(\boldsymbol{y}_t|\boldsymbol{x}_t)\rangle_{\tilde{p}(\boldsymbol{y})}\Big\rangle_q + H_q(X). \tag{58}$$

These terms pose different complexity challenges for the inference. The terms that only involve state variables $\boldsymbol{x}_t$ are relatively straightforward to compute. It suffices to rewrite (54) as

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) \propto \mathcal{G}(\boldsymbol{\xi}_t; \boldsymbol{0}, \Gamma), \tag{59}$$

where

$$\boldsymbol{\xi}_t = \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t+1}\end{bmatrix}, \qquad \Gamma^{-1} = \begin{bmatrix} A^\intercal Q^{-1} A & -A^\intercal Q^{-1} \\ -Q^{-1}A & Q^{-1}\end{bmatrix}, \tag{60}$$

and define

$$\lambda_t = \begin{bmatrix} \boldsymbol{m}_t \\ \boldsymbol{m}_{t+1}\end{bmatrix}, \qquad \Lambda_t = \begin{bmatrix} \Sigma_{t,t} & \Sigma_{t,t+1} \\ \Sigma_{t+1,t} & \Sigma_{t+1,t+1}\end{bmatrix}, \tag{61}$$

and

$$\hat{P}_{i,j} = \left\langle \boldsymbol{x}_i\boldsymbol{x}_j^\intercal\right\rangle_q = \Sigma_{i,j} + \boldsymbol{m}_i\boldsymbol{m}_j^\intercal, \quad 1\leqslant i,j\leqslant\tau. \tag{62}$$

From (8) and (53)-(56), the following equalities then hold, up to constants that do not depend on $\boldsymbol{m}$ and Σ,

$$\langle \ln p(\boldsymbol{x}_1)\rangle_q \propto -\frac{1}{2}\Big[||\boldsymbol{\mu}_0 - \boldsymbol{m}_1||_S^2 + \mathrm{tr}(S^{-1}\Sigma_{1,1})\Big]$$
$$= \boldsymbol{\mu}_0^\intercal S^{-1}\boldsymbol{m}_1 - \frac{1}{2}\mathrm{tr}(S^{-1}\hat{P}_{1,1}), \tag{63}$$

$$\langle \ln p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)\rangle_q \propto -\frac{1}{2}\Big[||\lambda_t||_\Gamma^2 + \mathrm{tr}(\Gamma^{-1}\Lambda_t)\Big]$$
$$= -\frac{1}{2}\mathrm{tr}(\Gamma^{-1}\Phi_t), \tag{64}$$

$$H_q(X) = \frac{1}{2}\ln|\Sigma|\,, \tag{65}$$

where

$$\Phi_t = \begin{bmatrix} \hat{P}_{t,t} & \hat{P}_{t,t+1} \\ \hat{P}_{t+1,t} & \hat{P}_{t+1,t+1}\end{bmatrix} = \Lambda_t + \lambda_t\lambda_t^\intercal\,. \tag{66}$$

These equations are similar to those of the LDS and can be computed efficiently. The main difficulty is the evaluation of the term

$$\Big\langle \langle \ln p(\boldsymbol{y}_t|\boldsymbol{x}_t)\rangle_{\tilde{p}(\boldsymbol{y})} \Big\rangle_q = \tag{67}$$
$$\sum_{k=1}^{K} \Big[\pi_{kt} \left\langle \ln \sigma(\omega_{kt}) \right\rangle_q + (1-\pi_{kt}) \left\langle \ln \sigma(-\omega_{kt}) \right\rangle_q \Big],$$

due to the non-linearity of the expectations $\langle \ln \sigma(\omega_{kt})\rangle_q$ and $\langle \ln \sigma(-\omega_{kt})\rangle_q$. Since ω (a linear projection of $\boldsymbol{x}$) is Gaussian, $\langle \ln \sigma(\omega)\rangle_q$ is bounded by

$$\langle \ln \sigma(\omega)\rangle_q \geqslant \ln \sigma(\langle \omega\rangle_q) - \frac{1}{8}\mathrm{var}(\omega), \tag{68}$$

which results from setting $\xi = 1/2$ in (A.10) of (Saul and Jordan, 2000). This leads to a new lower bound $\hat{\mathscr{L}}(\boldsymbol{\theta}, q)$ of $\langle \ln \mathcal{L}(\theta, y)\rangle_{\tilde{p}(y)}$ in (43)

$$\begin{aligned}\hat{\mathscr{L}}(\boldsymbol{\theta}, q) = &-\frac{1}{2}\Big\{ ||\boldsymbol{\mu}_0 - \boldsymbol{m}_1||_S^2 + \mathrm{tr}(S^{-1}\Sigma_{1,1}) \\ &+ \sum_{t=1}^{\tau-1} \mathrm{tr}(\Gamma^{-1}\Phi_t) + \frac{1}{4}\sum_t \mathrm{tr}(C\Sigma_{t,t}C^\intercal)\Big\} \\ &+ \sum_{t,k}\Big[\pi_{kt}\ln\sigma(\hat\omega_{kt}) + (1-\pi_{kt})\ln\sigma(-\hat\omega_{kt})\Big] \\ &+ \frac{1}{2}\ln|\Sigma| + \mathrm{const},\end{aligned} \tag{69}$$

where $\hat\omega_{kt} = \langle \omega_{kt}\rangle_q = C_{k\cdot}\boldsymbol{m}_t + u_k$.

The variational distribution $q^*(\boldsymbol{x})$ is the solution of

$$\{\boldsymbol{m}^*, \Sigma^*\} = \underset{\{\boldsymbol{m},\Sigma\}\in\mathbb{R}^{L\tau}\times\mathcal{S}_{++}^{L\tau}}{\arg\max} \hat{\mathscr{L}}(\boldsymbol{\theta}, q). \tag{70}$$

This is a *convex optimization* problem, since all terms of $\hat{\mathscr{L}}(\boldsymbol{\theta}, q)$, depend on either Σ or $\boldsymbol{m}$ separately (not on both), have the convex domain $(\boldsymbol{m}, \Sigma) \in \mathbb{R}^{L\tau}\times\mathcal{S}_{++}^{L\tau}$ and are concave - either a) linear functions, b) quadratic functions of negative definite coefficient matrices, c) negative log-sum-exp functions, or d) log determinant of Σ. Furthermore, (70) can be factorized into

$$\begin{aligned}\{\boldsymbol{m}^*, \Sigma^*\} &= \underset{\{\boldsymbol{m},\Sigma\}\in\mathbb{R}^{L\tau}\times\mathcal{S}_{++}^{L\tau}}{\arg\max} \hat{\mathscr{L}}(\boldsymbol{\theta}, q) \\ &= \Big\{ \underset{\boldsymbol{m}\in\mathbb{R}^{L\tau}}{\arg\max}\hat{\mathscr{L}}(\boldsymbol{\theta}, q), \underset{\Sigma\in\mathcal{S}_{++}^{L\tau}}{\arg\max}\hat{\mathscr{L}}(\boldsymbol{\theta}, q)\Big\}.\end{aligned}$$

Consolidating the terms containing Σ,

$$\begin{aligned}\Sigma^* = \underset{\Sigma}{\arg\max}\ &\ln|\Sigma| - \mathrm{tr}(W\Sigma), \\ &s.t.\ \Sigma \in \mathcal{S}_{++}^{L\tau},\end{aligned} \tag{71}$$

where $W \in \mathcal{S}_{++}^{L\tau}$ is a positive-definite matrix such that

$$W_{i,j} = \begin{cases} A^\intercal Q^{-1}A + S^{-1} + \frac{1}{4}C^\intercal C, & i=j=1, \\ A^\intercal Q^{-1}A + Q^{-1} + \frac{1}{4}C^\intercal C, & 1<i=j<\tau, \\ Q^{-1} + \frac{1}{4}C^\intercal C, & i=j=\tau, \\ -Q^{-1}A, & i=j+1, \\ -A^\intercal Q^{-1}, & i=j-1, \\ 0, & \text{otherwise}, \end{cases}$$

with $W_{i,j} \in \mathbb{R}^{L\times L}$ as block i, j of W. In Appendix B we show that this has optimal solution

$$\Sigma^* = W^{-1}. \tag{72}$$

While (72) is conceptually straightforward, the inversion of the matrix W can be too expensive for long video sequences (large τ). In sections C.1 and C.2 of Appendix C, we provide an alternative and more efficient procedure, based on the popular *Kalman smoothing filter* (Roweis and Ghahramani, 1999), to compute the parameters $\Sigma_{t,t}^*$ and $\Sigma_{t,t+1}^*$ needed to evaluate $\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)$,

The optimal variational mean parameter $\boldsymbol{m}^*$ has no closed form solution, due to the log-sigmoid terms of (69). In Appendix C.3, we discuss a numerical procedure for determining the stationary point of $\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)$. Since the problem is convex, this suffices to guarantee a global optimum.

6.3 The VLAD for Attribute Dynamics

The VLAD for attribute dynamics (VLADAD) approximates the Fisher score of the BDS by the derivatives of the variational lower bound of (69) with respect to the model parameters. In Appendix D we show that, given attribute sequence $\boldsymbol{\pi}$ and BDS $\boldsymbol{\theta} = \{S^{-1}, \boldsymbol{\mu}, A, Q^{-1}, C, \boldsymbol{u}\}$, [5] $\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)$ has derivatives

$$\frac{\partial}{\partial S^{-1}}\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*) = \frac{1}{2}\Big(S + \boldsymbol{\mu}\boldsymbol{m}_1^{*T} + \boldsymbol{m}_1^*\boldsymbol{\mu}^\intercal - \hat{P}_{1,1}^* - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal\Big), \tag{73}$$

$$\frac{\partial}{\partial \boldsymbol{\mu}}\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*) = S^{-1}(\boldsymbol{m}_1^* - \boldsymbol{\mu}), \tag{74}$$

$$\frac{\partial}{\partial A}\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*) = Q^{-1}(\Psi - A\phi), \tag{75}$$

$$\frac{\partial}{\partial Q^{-1}}\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*) = \frac{1}{2}\Big[\Psi A^\intercal + A\Psi^\intercal - A\phi A^\intercal - \varphi + (\tau-1)Q\Big], \tag{76}$$

[5] For simplicity, we consider the precision matrices S^{-1} and Q^{-1} instead of the covariances S, Q in the computation of Fisher scores.

$$\frac{\partial}{\partial \tilde{C}} \hat{\mathscr{L}}(\boldsymbol{\theta}, q^*) = \\ -\frac{1}{4}\left\{\tilde{C}\tilde{\Upsilon} + \sum_{t=1}^{\tau} \begin{bmatrix} \sigma(\tilde{C}_1.\boldsymbol{b}_t) - \pi_{1t} \\ \vdots \\ \sigma(\tilde{C}_K.\boldsymbol{b}_t) - \pi_{Kt} \end{bmatrix} \boldsymbol{b}_t^{\intercal}\right\}, \tag{77}$$

where

$$\varphi = \sum_{t=2}^{\tau} \hat{P}^*_{t,t}, \ \phi = \sum_{t=2}^{\tau} \hat{P}^*_{t-1,t-1}, \ \Psi = \sum_{t=2}^{\tau} \hat{P}^*_{t,t-1};$$

and $\tilde{C} = [C, \boldsymbol{u}]$, $\tilde{C}_k.$ is the k-th row of $\tilde{C}$,

$$\tilde{\Upsilon} = \begin{pmatrix} \sum_{t=1}^{\tau} \Sigma^*_{t,t} & 0 \\ 0 & 0 \end{pmatrix}, \ \boldsymbol{b}_t = \begin{pmatrix} \boldsymbol{m}^*_t \\ 1 \end{pmatrix}.$$

The VLADAD is then computed by 1) concatenating (73)-(77), and 2) aggregating over all attribute sequences extracted from a query video sequence. To improve discrimination, we apply a power-normalization and then $L2$-normalize the VLADAD feature vector, as suggested in (Jegou et al, 2012).

7 Experiments

In this section, we discuss experiments designed to evaluate the performance of the proposed BDS, BoWAD, and VLADAD. Three benchmarks from various perspectives are adopted to assess the behavior of these approaches: the *Weizmann Complex Activity* is a synthetic benchmark with comprehensive simulated challenges; *Olympic Sports* contains weakly cropped and aligned complex sport sequences; and *Multimedia Event Detection* features high level events with instances from open-source repositories.

7.1 Attribute classifiers

The VLADAD can be computed for any implementation of attribute classifiers. Since the goal was not attribute detection *per se*, we used two popular methods to produce attribute sequences. The first attribute classifier extracted space-time interest points (STIP) of (Laptev, 2005) and computed at each interest point a descriptor combining a histogram of oriented gradients (HoG) and a histogram of optical flow (HoF). The second classifier was based on the improved trajectory feature (ITF) of (Wang and Schmid, 2013), using a descriptor composed of HoG, HoF, frame-wise trajectory (FWT), and motion boundary histogram (MBH), which has been shown to achieve state-of-the-art performance in action recognition even superior than features by deep learning (Karpathy et al, 2014; Peng et al, 2014; Simonyan and Zisserman, 2014; Wang et al, 2015). All features were extracted with the binary or source code provided by its authors. [6]

In all experiments, attribute detection was based on the BoVW. For each descriptor, a codebook of size V was learned by k-means, over the entire training set, and used to quantize features. Different ITF descriptors were processed separately and merged by averaging kernel matrices during prediction. The attribute annotations of (Liu et al, 2011) were used for Weizmann and Olympic Sports and those of (Bhattacharya, 2013) for MED. Appendix F provides details on attribute definitions and annotations. On Weizmann, attribute detectors were implemented with a linear SVM, using LIBSVM (Chang and Lin, 2011) with probability outputs. However, we found this to have scalability problems for the larger Olympics and MED datasets. On these datasets attribute classifiers were logistic regressors, implemented with LIBLINEAR (Fan et al, 2008). To maximize attribute detection accuracy, while retaining the efficiency of linear classification, we used an additive kernel mapping of the histogram intersection kernel (HIK), as suggested in (Vedaldi and Zisserman, 2012). The attribute trajectory $\{\boldsymbol{\pi}_t\}$ of a video sequence was computed with a sliding window, where attribute detectors predicted attribute scores at each window anchoring position. An holistic attribute vector, encoding the presence of attributes in the entire video sequence, was also constructed by max-pooling $\{\boldsymbol{\pi}_t\}$ over time.

7.2 Weizmann Complex Activity

The first set of experiments aimed to systematically compare the ability of different models to capture the dynamics of attribute sequences. A non-trivial difficulty of such a study is the need for datasets with classes that 1) differ only in terms of attributes dynamics, and 2) enable a quantification of these differences. It is critical that such datasets do not include discriminant information beyond attribute dynamics, such as discriminant scene backgrounds, objects, or scene durations. Unfortunately, these conditions are not met by existing action datasets. For example, the “making a sandwich” activity of the MED dataset is the only one to include the “sandwich” object. This enables the use of object recognition as a proxy for action recognition, an alternative that would not be viable if the dataset also contained an “eating a sandwich” activity. To avoid these problems,

[6] Binary for STIP available at http://www.di.ens.fr/~laptev/download; source code for ITF available at http://lear.inrialpes.fr/~wang/download.

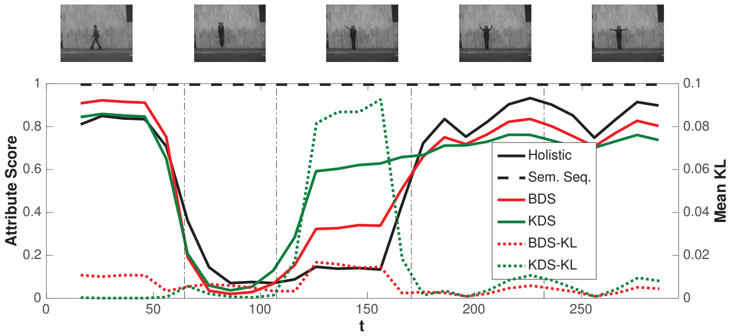


Fig. 7: Top: key frames from activity sequence class "walk- pjump-wave1-wave2-wave2" in Syn-4/5/6. Bottom: score of "two-arms-motion" attribute. True scores in black, and scores sampled from BDS (red) and KDS (blue). Also shown is the KL-divergence between sampled and true scores, for both models.

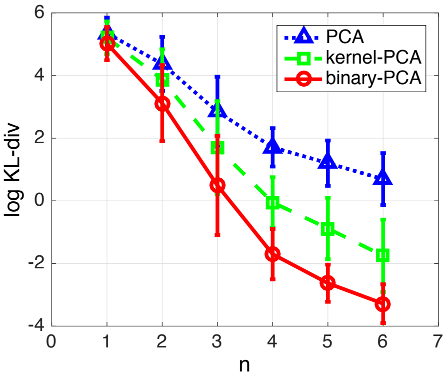


Fig. 8: Log KL-divergence between original and reconstructed attribute scores, *v.s.* number of PCA components n, on Syn-4/5/6 for PCA, KPCA, and binary PCA.

we assembled a synthetic dataset of complex sequences, which were synthesized from the atomic actions of the popular Weizmann dataset (Gorelick et al, 2007).

Weizmann contains 10 *atomic* action classes (*e.g.*, skipping, walking) performed by 9 people and was annotated with 30 low-level attributes (*e.g.*, "one-arm-motion") by (Liu et al, 2011). Attribute sequences were computed over 30-frame sliding video windows with 10-frame stride. STIP features were used with a 1000-word vocabulary for low-level descriptor quantization. The availability of attribute ground truth for all atomic actions enables learning of clean attribute models. Hence, performance variations can be attributed to the quality of the attribute-based inference of the different approaches.

Three subsets of synthetic sequences were created by concatenating Weizmann actions (see Appendix E for some examples). These subsets vary in the variability and complexity of temporal structure of their video sequences. They target the study of different hypotheses regarding the role of dynamics in action recognition. The first, denoted "Syn-4/5/6" evaluates the ability of different models to capture dynamics of varying complexity, when all video segments are informative of the action class, *i.e.*, when the dynamics have no noise. The remaining two evaluate robustness to "noisy dynamics". "Syn20 × 1" consists of actions of homogeneous dynamics, which are buried in additional video segments of dynamics uncharacteristic of the action class. "Syn10×2" consists of discontinuous actions of homogeneous dynamics, which are interleaved with segments of "noisy dynamics".

7.2.1 Complex Dynamics

In the first subset, "Syn-4/5/6", a sequence of *degree* n ($n = 4, 5, 6$) is composed of n atomic actions, performed by the same person. The row of images at the top of Figure 7 presents keyframes of an activity sequence of degree 5, composed by the atomic actions "walk", "pjump", "wave1", "wave2", and "wave2". The black curve (labeled "Sem. Seq") in the plot at the bottom of the figure shows the score of the "two-arms-motion" attribute over time. 40 activity categories were defined per degree n (total of 120 activity categories), and the dataset was assembled per category, containing one activity sequence per person (9 people, 1080 sequences in total). Overall, the activity sequences differ in the number, category, and temporal order of atomic actions.

We started by comparing the binary PCA that underlies the BDS to the PCA and KPCA decompositions of the LDS and KDS. In all cases, a set of attribute score vectors $\{\boldsymbol{\pi}_t\}$ was projected into the low-dimensional PCA subspace, the reconstructed score vectors $\{\hat{\boldsymbol{\pi}}_t\}$ were computed and the KL divergence between $B(\boldsymbol{y}, \boldsymbol{\pi}_t)$ and $B(\boldsymbol{y}, \hat{\boldsymbol{\pi}}_t)$ was measured. The logit kernel $K(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$ $= \sigma^{-1}(\boldsymbol{\pi}_1)^\intercal \sigma^{-1}(\boldsymbol{\pi}_2)$, where $\sigma^{-1}(\cdot)$ is the element-wise logit function, was used for KPCA. Fig. 8 shows the average log-KL divergence, over the entire dataset, as a function of the number of PCA components used in the reconstruction. Binary PCA outperformed both PCA and KPCA. The improvements over KPCA are particularly interesting, since the latter uses the logistic transformation that distinguishes binary PCA from PCA. This is explained by the Euclidean similarity measure that underlies the assumption of Gaussian noise in KPCA, as discussed in Section 4.6.

Table 1: Accuracy on Syn-4/5/6.

method		accuracy
BoVW (Laptev et al, 2008)	(x1y1t1)	57.8%
	(x1y1t3)	78.8%
	(x1y1t6)	92.5%
holistic attribute		72.6%
DTM (Blei and Lafferty, 2006)		84.6%
ToT (Wang and McCallum, 2006)		88.2%
KDS (Chaudhry et al, 2009)		90.2%
BDS		94.8%

To gain some more insight on the different models, a KDS and a BDS were learned from the 30 dimensional attribute score vectors of the activity sequence in Figure 7. A new set of attribute score vectors were then sampled from each model. The evolution of the scores sampled for the "two-arms-motion" attribute are shown in the figure (in red/blue for BDS/KDS). Note how the scores sampled from the BDS approximate the original attribute scores better than those sampled from the KDS. This was quantified by computing the KL-divergences between the original attribute scores and those sampled from the two models, which are also shown in the figure.

We next evaluated the benefits of different representations of dynamics for activity recognition. Recognition rates were obtained with a 9-fold leave-one-out-cross-validation (LOOCV), where, per trial, the activities of one subject were used as test set and those of the remaining 8 as training set. We compared the performance of classifiers based on the KDS and BDS to those of a BoVW classifier with temporal pyramid (TP) matching (Laptev et al, 2008), a holistic attribute classifier that ignores attribute dynamics, the dynamic topic model (DTM) (Blei and Lafferty, 2006) and the topic over time (ToT) model (Wang and McCallum, 2006) from the text literature. For the latter, topics were equated to the activity attributes and learned with supervision (using the SVMs for attribute detection). Unsupervised versions of the topic models had worse performance and are omitted. Classification was performed with Bayes' rule for topic models, and a nearest-neighbor classifier for the remaining methods. BDS distances were measured with (27), while for the KDS we adopted the logit kernel. The dimension of the BDS state space was 5. The $\mathcal{X}^2$ distance was used for all BoVW and holistic attribute classifiers. In an attempt to match the pooling mechanism of temporal pyramid matching to the structure of the synthetic Weizmann sequences, we considered a variant with 6 temporal bins. This is denoted BoVW-x1y1t6.

Table 2: Accuracy on Syn20×1 and Syn10×2.

method		Syn20×1	Syn10×2
BoVW (Laptev et al, 2008)	(x1y1t1)	23.3%	28.9%
	(x1y1t3)	36.7%	31.1%
	(x1y1t6)	55.6%	24.4%
holistic attribute		17.8%	16.7%
DTM (Blei and Lafferty, 2006)		49.3%	46.5%
ToT (Wang and McCallum, 2006)		57.2%	55.9%
KDS (Chaudhry et al, 2009)		61.6%	63.1%
BDS		64.4%	65.6%
BoWAD	(BMC)	100%	100%
	(MDS-kM)	100%	98.9%
VLADAD	(BMC)	100%	100%
	(MDS-kM)	100%	100%

The accuracy of all classifiers is reported in Table 1. BDS achieved the best performance, followed by BoVW-x1y1t6, KDS, the dynamic topic models, and BoVW-x1y1t1 and holistic attribute. Note the large difference between the holistic attribute and the best dynamic model ($\approx 22\%$). This shows that while attributes are important (14.8% improvement over BoVW without temporal pooling), they are not the whole story. Problems involving *fine-grained* activity classification, *i.e.*, discrimination between activities composed of similar actions executed in different sequence, requires modeling of attribute dynamics. This is reflected by both the improvement of BoVW with $x1y1t3$ and $x1y1t6$ temporal pyramids over naive BoVW, and that of models of attribute dynamics over the holistic attribute vector. Among the dynamic models, the BDS outperformed the KDS, the topic models DTM and ToT, and BoVW with pyramids $x1y1t3/t6$. It is also worth noting the sensitivity of pyramid matching to the number of temporal bins, with performance varying between 57.8% (x1y1t1) and 92.5% (x1y1t6).

7.2.2 Noisy dynamics

The remaining two datasets evaluated the robustness of the different methods to noise, poor segmentation, and alignment. The second dataset, "Syn20×1" was composed of activity classes of large variability. Each activity was defined as a sequence of 20 *consecutive* atomic actions. This sequence was inserted at a *random* temporal location of a larger sequence of 40 atomic actions. The remaining 20 actions in the larger sequence were randomly selected from Weizmann. The third dataset, "Syn 10×2", tested the detection of *discontinuous* activities. Each activity was defined by two subsequences, each with 10 consecutive atomic actions. The two subsequences were randomly inserted at non-overlapping

Table 3: Mean average precisions on Olympic Sports.

method		w/o LA fusion		w/ LA fusion	
		STIP	ITF	STIP	ITF
BoVW	(x1y1t1)	59.0%	83.7%	-	-
(Laptev et al, 2008)	(x1y1t3)	53.2%	81.6%	-	-
DMS (Niebles et al, 2010)		62.5%	-	-	-
holistic attribute		62.6%	82.1%	64.2%	84.9%
VD-HMM (Tang et al, 2012)		66.8%	-	-	-
HMM-FV (Sun and Nevatia, 2013)		65.3%	84.7%	66.4%	86.7%
CTR (Bhattacharya et al, 2014)		64.9%	85.5%	67.1%	87.3%
BDS		67.8%	86.1%	68.7%	88.6%
BoWAD	(BMC)	73.5%	90.3%	74.9%	91.2%
	(MDS-kM)	71.2%	88.2%	72.6%	89.8%
VLADAD	(BMC)	**76.9%**	**91.7%**	**77.2%**	**93.1%**
	(MDS-kM)	71.7%	90.6%	73.4%	91.4%

Table 4: Performance on Olympic Sports.

method	mAP
Todorovic (2012)	82.9%
Jain et al (2013b)	83.2%
Jain et al (2013a)	85.3%
Li et al (2013a)	84.5%
Wang and Schmid (2013)	91.1% [a]
Jones and Shao (2014)	74.6%
Ni et al (2015)	92.3%
Lan et al (2015)	92.9%
VLADAD	**93.1%**

[a] Result achieved with 1) feature points pruned by human detection, and 2) Fisher vector encoding of low-level features. Without these enhancements, the performance of (Wang and Schmid, 2013) is 83.3%.

locations of the larger (40 atomic actions) sequence. For both sets, 20 activities were synthesized for each of 9 subjects, producing 180 sequences per set.

In addition to the classifiers of Table 1, both the BoWAD and VLADAD were evaluated on these datasets. For both, short-term attribute sequences consisted of attribute vectors from 12 consecutive windows. The dimension of the BDS state space was again 5. WAD dictionaries were learned with both BMC and the MDS-kM algorithm of (Ravichandran et al, 2012). One-versus-all SVMs were used for BoVW and BoWAD classification, using a χ^2 kernel. VLADAD was implemented with a linear kernel, KDS and BDS used the kernel $K(\boldsymbol{\Omega}_a, \boldsymbol{\Omega}_b) = \exp(-\frac{1}{\gamma} d^2(\boldsymbol{\Omega}_a, \boldsymbol{\Omega}_b))$ where d is the distance used in Syn-4/5/6. These kernels achieved the best performance for each of the methods in our preliminary experiments.

Table 2 summarizes the performance of the different methods. Both BoVW and the holistic attribute vector performed poorly. Note, in particular, how BoVW-x1y1t6 now underperformed the two other implementations of temporal pyramid matching. This highlights the difficulty of designing universal pooling schemes, that can withstand significant intra class variability. This problem also affected the dynamics models, which performed substantially worse than in Table 1. While the BDS significantly outperformed the other methods, its performance was still lackluster. This is explained by the underlying assumption of a single dynamic process, a severe mismatch on Syn20×1 and Syn10×3, where the activities of interest are 1) not temporally aligned and 2) immersed in irrelevant video content. It is thus not surprising that the BoWAD and VLADAD achieved substantially better performance on these datasets, reaching perfect classification. With respect to BoWAD clustering, both strategies achieved excellent results, with BMC performing slightly better than MDS-kM. Overall, these results demonstrate the robustness of the proposed BoWAD and VLADAD representations to intra-class variation and noise.

7.3 Olympic Sports

The second set of experiments was performed on Olympic Sports (Niebles et al, 2010). This contains YouTube videos of 16 sport activities, with a total of 783 sequences. Some activities are sequences of atomic actions, whose temporal structure is critical for discrimination from other classes (*e.g.*, “clean and jerk” *v.s.* “snatch”, and “long-jump” *v.s.* “triple-jump”). Since the attribute labels of (Liu et al, 2011) are only available for whole sequences, the attribute classifiers are much noisier than in the previous experiment, degrading the quality of attribute models. We followed the train-test split proposed by (Niebles et al, 2010) and used per-category average precision (AP) and mean AP (mAP) to measure recognition performance. In all cases, low-level feature quantization was based on 4000-word codebooks, learned with k-means. Attribute sequences were computed with a 30-frame sliding window, implemented with a stride of 4 frames.

The proposed approaches were compared to BoVW-TP, the decomposable motion segments model (DMS) of (Niebles et al, 2010), the hidden Markov model with latent states of variable duration (VD-HMM) (Tang et al, 2012), the holistic attribute, and two recent approaches that also model attribute dynamics: the HMM fisher vector (HMM-FV) of (Sun and Nevatia, 2013) and the combined temporal representation (CTR) of (Bhattacharya et al, 2014). Classification was performed with SVMs using a χ^2 or Jensen-Shannon kernel for

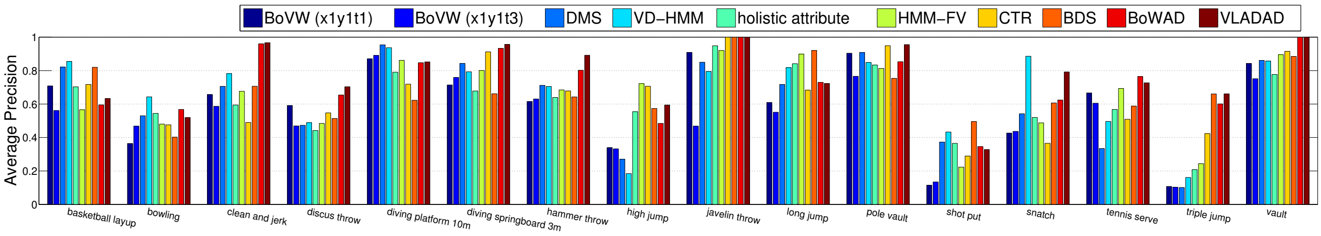


Fig. 9: Average precisions on Olympic Sports with STIP as the low-level feature.

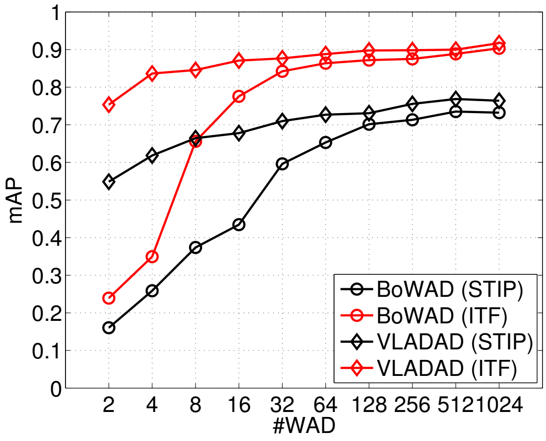


Fig. 10: Mean average precision (mAP) *v.s.* size of WAD dictionary on Olympic Sports.

histogram-based methods (BoVW, holistic attribute, BoWAD); SVMs using a radial basis function (RBF) kernel $K_\alpha(i,j) = \exp(-\frac{1}{\alpha}d^2(i,j))$ for HMM-FV and CTR; a nearest neighbor classifier or SVM using the RBF kernel for BDS; and a linear SVM for VLADAD. For each method, the best classifier parameters were chosen by 4-fold cross-validation on the training set. The number of PCA components L of the BDS was selected from $\{2, 4, 6, 8\}$, and the length τ of the attribute sequences of BoWAD and VLADAD from $\{4, 6, 8, 10, 12, 16\}$ by cross-validation on the training set.

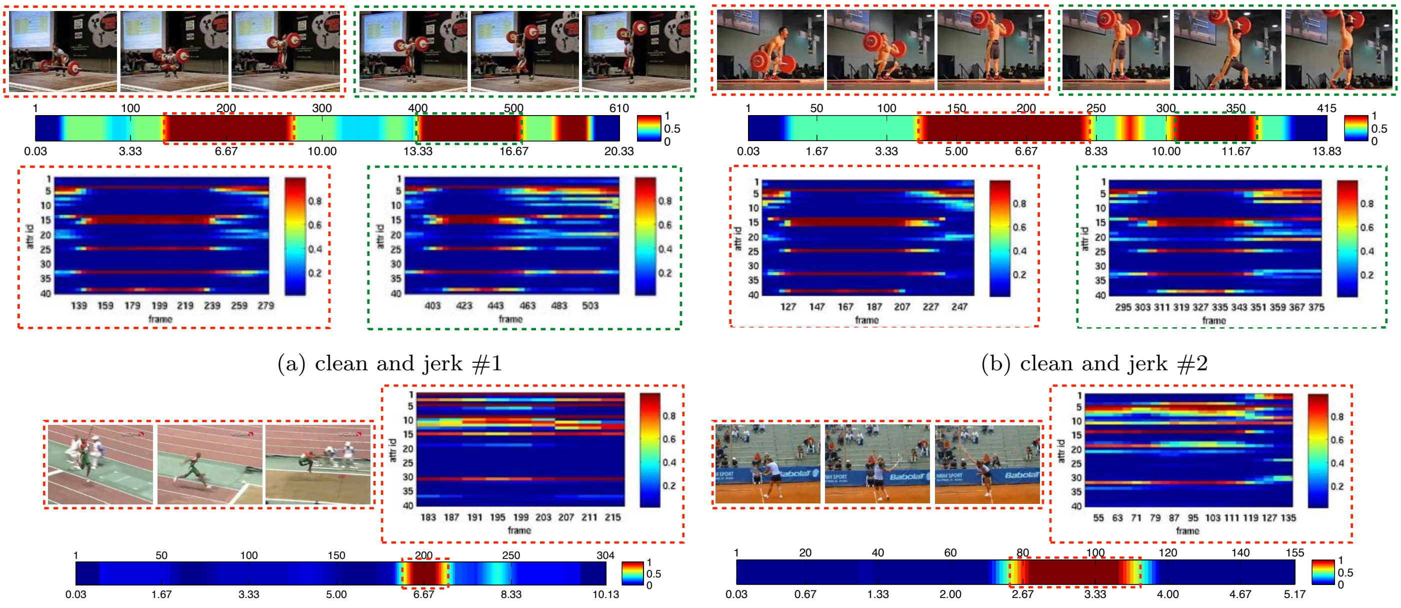
The performance of the different approaches is summarized in Table 3[7]. Several conclusions can be drawn. First, all models benefit strongly from the ITF features. The increased performance of BDS, BoWAD, and VLADAD with these features suggests that a more discriminant set of low-level features, and thus cleaner attributes, can significantly simplify the problem of modeling of attribute dynamics.

Second, the BDS again outperforms all other models. The gains are larger over methods that do not account for dynamics (*e.g.*, the holistic attribute vector) but substantial even over the alternative models of attribute dynamics, such as HMM-FV or CTR. This is likely due to the richer characterization of the hidden state space by the BDS and its modeling of low-dimensional attribute subspaces. An interesting observation is that BoVW-x1y1t3 underperforms the vanilla BoVW significantly, reflecting the fact that its rigid temporal cells with fixed temporal anchor points 1) are coarse for capturing finer structure within each cell, and 2) cannot adapt to intra-class variation. This vulnerability of BoVW with augmented "rigidity" to overfitting is also confirmed by other works in literature (Lan et al, 2014).

Third, the BDS gains are smaller than in Weizmann. This is due, in part, to the increased difficulty of modeling dynamics because annotations are noisy and, in part, to the nature of the dataset. While Weizmann requires fine-grained temporal discrimination for most classes, this is not the case in Olympic. For example, the holistic attribute vector suffices to discriminate classes that are very distinctive, *e.g.*, that have *unique* motion. An example is "diving platform 10m," which can be singled out by its distinctive patterns of fast downward motion. This is visible in the per-category average-precision plot of Fig. 9, where the holistic attribute vector performs very well for this class. On the other hand, finer grained temporal analysis is required to distinguish between similar classes, *e.g.*, "long-jump" *v.s.*"triple jump", or "clean and jerk" *v.s.*"snatch". Fig. 9 clearly shows that these classes 1) pose a greater challenge to previous methods, and 2) lead to the largest gains by the BDS, BoWAD, and VLADAD.

Fourth, while the BDS performs quite well for classes with reasonably well segmented and aligned sequences (*e.g.*, "long jump"), the assumption of a single dynamic process again limits its performance for categories with larger variability (*e.g.*, "snatch", "clean and jerk", "tennis serve", *etc*). Both BoWAD and VLADAD perform better in this case, improving BDS performance by 4% to 9% overall. Fig. 9 shows that this improvement is particularly significant for categories, such as "clean

[7] Note that the version of Olympic Sports used in (Niebles et al, 2010) is different from that released publicly. DMS performance on the latter was reported in (Tang et al, 2012).



(a) clean and jerk #1

(b) clean and jerk #2

(c) triple jump

(d) tennis serve

Fig. 11: Recounting by the BoWAD on Olympic sports (best viewed in color). The normalized score, for activity recognition, of each video segment is shown as a bar (time in seconds displayed at the bottom, frame id at the top). As shown in the color key, red corresponds to a score of 1 (most relevant), blue to a score of 0 (less relevant). The dashed lines identify the most significant events. Associated key-frames are shown at the top, corresponding attribute sequences at the bottom. Same setting applies to all recounting illustrations.

and jerk” and “tennis serve”, whose discriminant events are scattered throughout the video sequence.

Fifth, regarding encoding schemes there is now a clear gap between BoWAD (73.5% for STIP, 90.3% for ITF) and VLADAD (76.9% for STIP, 91.7% for ITF). This confirms many previous observations for the effectiveness of Fisher scores in image and video classification. Fig. 10 shows that the VLADAD gains hold across a substantial range of WAD codebook size. Note that a 16-word VLADAD codebook already has mAP (around 87%) superior to most methods in Table 4. Similarly, we observed a consistent advantage of BMC over MDS-kM clustering, with differences of 1% to 5% in mAP (see Table 3).

Sixth, Fig. 9 shows that even methods with low overall performance, *e.g.*, the holistic attribute vector, can have good performance for some classes. This suggests that there is some complementarity in the different video representations, and it may be beneficial to combine them (Snoek et al, 2005; Tamrakar et al, 2012; Ye et al, 2012). We have investigated this by combining representations based on low-level features, holistic attributes, and dynamic modeling, using the late fusion scheme of (Tamrakar et al, 2012), which uses the geometric mean of scores of different classifiers as the final prediction score. The combination of multiple representations, denoted “LA fusion” in Table 3), does not change the conclusions above. Again, the BDS outperforms previous models of temporal structure, based on either low-level motion (DMS and VD-HMM) or attributes (HMM-FV and CTR), the BoWAD outperforms the BDS, and the VLADAD has top performance.

Seventh, all methods benefit from late fusion. This confirms that some discriminant information might be discarded by attribute modeling (gains by inclusion of low-level features) and holistic modeling can sometimes be useful. However, the effect is small, with a gain less than 1% for most the best performing methods.

Finally, Table 4 compares the VLADAD-BMC with ITF features and late fusion to previous approaches in the literature. The proposed representation achieves state-of-the-art performance on this dataset, surpassing the previous best results by (Wang and Schmid, 2013; Ni et al, 2015; Lan et al, 2015). Note that all these three benchmarks are based on ITF encoded with Fisher vector, which is a stronger baseline than ours (ITF with vanilla BoVW). This enhancement could be incorporated into our attribute detectors, potentially leading to even better performance.

Table 5: Event list for MED11

ID	Event Name	ID	Event Name
E001	attempt a board trick	E009	get a vehicle unstuck
E002	feed an animal	E010	groom an animal
E003	land a fish	E011	make a sandwich
E004	wedding ceremony	E012	parade
E005	work on a wood project	E013	parkour
E006	birthday party	E014	repair an appliance
E007	change a vehicle tyre	E015	work on a sewing project
E008	flash mob gathering		

7.4 Recounting

An interesting property of the BoWAD is that it can be easily combined with "recounting" procedures to support semantic video segmentation, summarization, and activity identification. This follows from the fact that the contribution of a particular WAD to the score of an activity classifier can be seen as a measure of the importance of the corresponding pattern of attribute dynamics for the detection of the target activity. We used the recounting procedure of (Yu et al, 2012), quantifying the significance of a video segment (for event detection) by the weighted sum of the similarities between the corresponding BoWAD histogram bin and those of the SVM support vectors. More specifically, let x be the BoWAD histogram and consider a prediction rule based on an additive kernel, *e.g.*, an SVM with HIK. In this case,

$$h(x) = \sum\nolimits_i \alpha_i g(x, z^{(i)}) + c, \tag{78}$$

where $z^{(i)}$ is the i-th support vector, α_i the corresponding SVM weight, c a constant, and $g(x, z^{(i)}) = \sum_j g_j(x_j, z^{(i)})$ measures the similarity between z_i and x. The prediction rule then can be rewritten as

$$h(x) = \sum\nolimits_{j,i} \alpha_i g_j(x_j, z^{(i)}) + c = \sum\nolimits_j h_j(x_j) + c, \tag{79}$$

where $h_j(x_j) = \sum_i \alpha_i g_j(x_j, z^{(i)})$ is the contribution of histogram bin x_i to the classification score of the BoWAD histogram. Note that, unlike the holistic attributes of (Yu et al, 2012), for which temporal localization intractable, each video segment is associated with a WAD in the BoWAD, which corresponds to a short-term pattern of activity. This allows the quantification of the contribution of the video segment to event detection by $h_j(x_j)$, where x_j is the bin of the corresponding WAD. This enables a precise characterization of the temporal duration and anchor points of different event evidence.

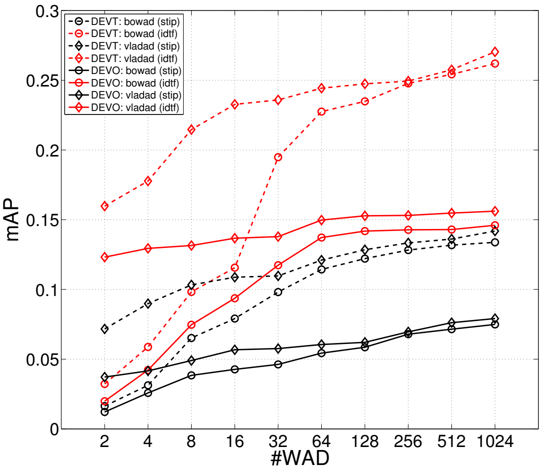


Fig. 12: Mean average precision (mAP) *v.s.* size of WAD dictionary on MED11.

Four examples are illustrated in Fig. 11. In both instances of "clean and jerk", the BoWAD discovers the two signature motion of "lifting barbell to chest level" and "lifting barbell over head". Note the variation in temporal location and duration of these events in the two sequences. On the other hand, the signature events discovered for "triple jump" and "tennis serve," are "large step forward followed by jump", and "toss ball into the air followed by hit," respectively. These results illustrate the robustness of the BoWAD to video uninformative of the target activity, and its ability to zoom in on the discriminant events. This is critical for accurate activity recognition from realistic video.

7.5 TRECVID-MED11

The third set of experiments used the 2011 TRECVID multimedia event detection (MED11) open source dataset (Over et al, 2011). This is one of the most challenging datasets for activity or event recognition due to 1) the vaguely defined high-level event categories (*e.g.*, "birthday party"); 2) the large intra-class variation in terms of event composition (*e.g.*, temporal duration, organization), stage setting, illumination, cutting, resolution, *etc*; 3) large negative samples, and so forth. We followed the protocol suggested by the TRECVID evaluation guidelines for performance evaluation. Specifically, the event collection (EC) set was used for training. EC contains 2,392 training samples of 15 high-level events (see Table 5 for the full list), with 100-200 positive examples per event. Two evaluation sets, DEV-T and DEV-O, were used for testing. DEV-T has 10,723 samples (370 hours of video in total), approximately

Table 6: Mean average precisions (in %) on MED11.

method		DEVT				DEVO			
		w/o LA fusion		w/ LA fusion		w/o LA fusion		w/ LA fusion	
		STIP	ITF	STIP	ITF	STIP	ITF	STIP	ITF
random guess		0.98				0.37			
BoVW	(x1y1t1)	15.70	32.68	-	-	8.31	18.53	-	-
(Laptev et al, 2008)	(x1y1t3)	15.50	31.86	-	-	9.66	18.92	-	-
DMS (Niebles et al, 2010)		5.72	-	-	-	2.52	-	-	-
holistic attribute		10.62	25.03	16.31	33.42	4.93	12.45	8.93	19.67
VD-HMM (Tang et al, 2012)		11.25	-	-	-	4.77	-	-	-
HMM-FV (Sun and Nevatia, 2013)		8.15	21.82	16.50	33.77	4.49	11.64	9.52	20.08
CTR (Bhattacharya et al, 2014)		9.46	22.42	17.14	33.61	4.62	11.08	9.61	19.72
BDS		6.75	16.72	16.33	33.49	3.67	9.21	9.16	19.21
BoWAD	(BMC)	13.38	26.20	18.05	35.02	7.49	14.36	10.25	20.91
	(MDS-kM)	12.70	25.08	17.37	34.11	6.92	13.68	9.94	20.30
VLADAD	(BMC)	14.19	27.04	18.56	**35.40**	7.91	15.61	10.92	**21.84**
	(MDS-kM)	13.41	26.16	17.93	34.62	7.33	14.84	10.15	20.89

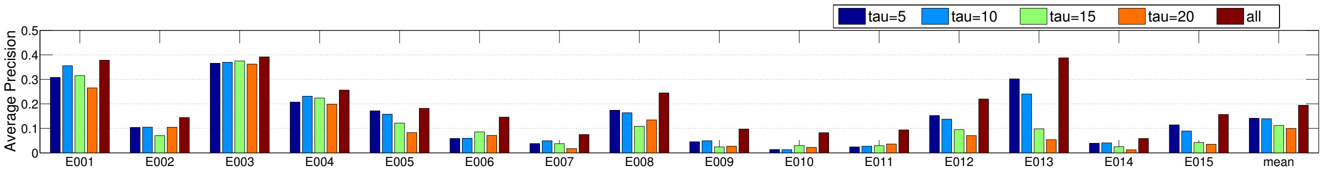


Fig. 13: Average precision of VLADAD using ITF for different segment lengths on MED11.

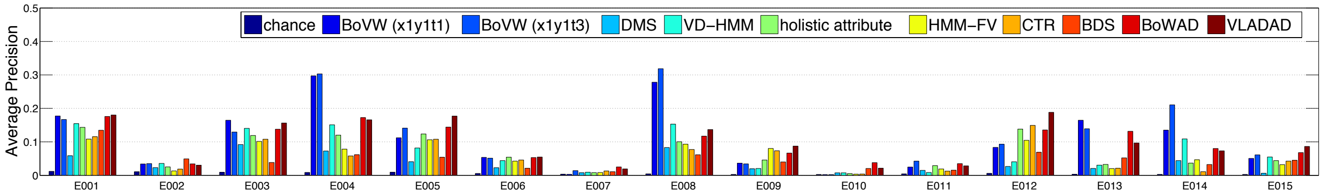


Fig. 14: Average precisions for different methods using STIP on MED11.

1% of which is from events 1-5 and the remaining 99% are negative samples; while DEV-O has 32,061 samples (1200 hours in total), with around 0.5% from events 6-15 and 99.5% negative samples.

Attribute classification was based on 103 attributes defined by (Bhattacharya, 2013). 8,000-word codebooks were learned with k-means for low-level feature quantization. Attribute scores were computed with a 180-frame sliding window and a 30-frame stride. All classifier settings were as in Section 7.3, with the exception of the length τ of attribute sequences for BoWAD and VLADAD, which was selected from $\{5, 10, 15, 20\}$, corresponding to roughly 5, 10, 15 and 20 seconds. To account for the variability of instances from the same event, both the BoWAD histograms and VLADAD were computed with different τ and concatenated into the feature used for event prediction.

Table 6 summarizes the event detection performance of the different methods. Most of these results are in line with those of the previous section. For example, the VLADAD again outperformed the BoWAD, especially for small codebook sizes. This is shown in greater detail in Fig. 12. Similarly, clustering with the BMC again outperformed MDS-kM. Finally, Fig. 13 shows the APs of VLADAD for different attribute sequence lengths τ. Not surprisingly, different lengths performed best for different events. For example, while in "parkour" (E013) the discriminant motion of "rush-jump-climb-land" takes about 5 seconds, in "land a fish" the distinctive motion of "pull-throw-catch" lasts between 5 and 20 seconds. Combing different attribute

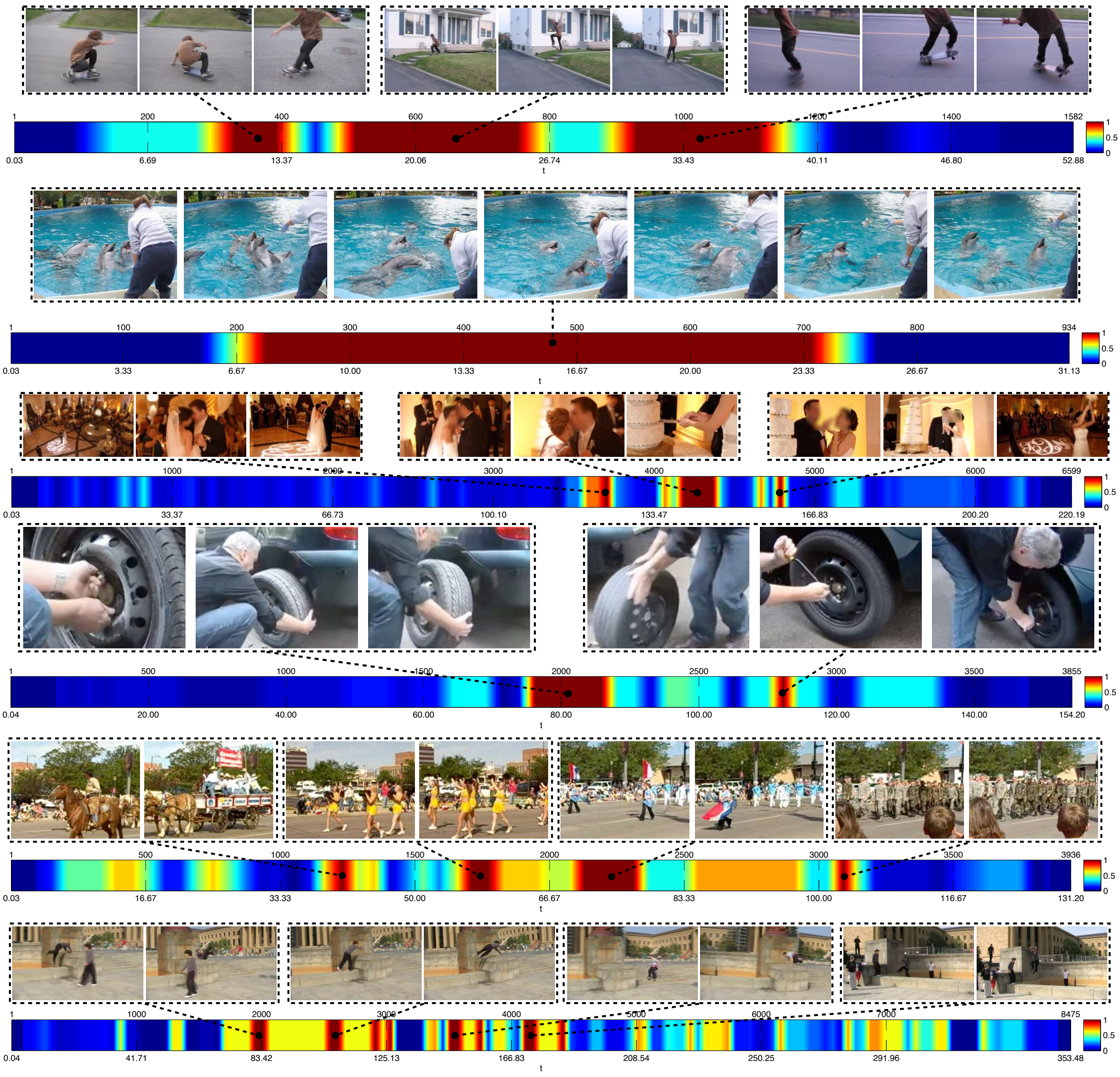


Fig. 15: Recounting by BoWAD on MED11 for sequences of "attempt a board trick", "feed an animal", "wedding ceremony", "change a vehicle tyre", "parade", and "parkour" (top to bottom). Snapshots from the most significant clips of each sequence are also shown.

sequence lengths achieved the best performance for all event classes.

However, there were also some significant differences. First, the previously proposed models of temporal structure, either for low-level features (DMS and VD-HMM) or attributes (BDS, HMM-FV, CTR), performed worse or, at most, on par with the holistic attribute vector. This can be justified by the complexity and variability of the MED events. The BDS was particularly affected by this problem, performing 1% − 5% worse than the other models of attribute dynamics. Together with Section 7.3, these results confirm that, while the BDS is a better model of dynamics for segmented and aligned video, it has difficulties for video containing multiple dynamic processes. The fact that the BoWAD and VLADAD outperform both the holistic attribute vector, and the previous models of low-level (DMS, VD-

HMM) and attribute (HMM-FV, CTR) dynamics shows that they effectively address this problem.

Second, and more surprising, attribute-based models underperformed the BoVW. This could be due to 1) noisy attribute classification, or 2) limited attribute vocabulary. Since, as shown in Fig. 14, attribute-based approaches handled some events better than the BoVW we believe that the latter is the main problem. In any case, since this shows that attribute representations capture information complementary to that of the BoW, the fusion of attribute models and the BoVW should lead to the best performance. Table 6 shows that this is indeed the case, as all attribute representations improve on the BoVW when combined with it by late fusion. In fact, when fused with BoVW and holistic attribute, the VLADAD achieves 21.84% mAP on MED11 DEV-O. In comparison to other benchmarks, this is substantially higher than the 15.69% of (Vahdat et al, 2013), 16.02% of (Lai et al, 2014), 15.35% of (Hajimirsadeghi et al, 2015), and comparable to 22.13% (best results for a single low-level feature) by (Xu et al, 2014).

Another important task in TRECVID is recounting of multimedia events, which we implemented as in Section 7.3. Several BoWAD recounting examples are illustrated in Fig. 15, again showing that modeling local signature behavior is sufficient for accurate detection of complex activities. Specifically, the BoWAD captures a somersault by a subject riding a skateboard in “attempt a board trick”, the action of throwing food to dolphins in “feeding an animal”, the scattered scenes of “dancing”, “cutting cake”, and “bouquet toss” in “wedding ceremony”, the marching crowd on “parade”, and so on. On the other hand, as shown in Fig. 16, recounting results also reveal two major reasons for detection false positives. The first is the existence of visual content (*e.g.*, motion) confusable with that of the target event. The top sequence of Fig. 16, a sequence of “attempt a board trick” where a bike rider performs somersaults similar to those executed by skateboard riders in the background, is an example of this problem. Similarly, the second sequence shows a false positive for “parkour,” where several athletes perform plyometric activities or other forms of training, which involve running, jumping over obstacles, and climbing. The second reason for false positives is the ambiguity of certain activities, which lead to inconsistent ground-truth on MED11. For example, the third and fourth sequences of Fig. 16 are labeled as background events for “groom an animal” and “parade,” respectively. However, the recounting results show that both sequences are indeed instances of these events.

8 Conclusion

In this work, we have proposed a novel representation for video, based on the modeling of action attribute dynamics. The core of this representation is the binary dynamic system (BDS), a joint model for attribute appearance and dynamics. This model was shown to be effective for video sequences that display a single activity, of homogeneous dynamics. To address the challenges of complex activity recognition, where video sequences can be composed of multiple atomic events or actions, the BDS was embedded in a BoVW-style representation, denoted the BoWAD. This is based on a BDS codebook, representing video as an histogram of assignments to BDSs that characterize temporally localized attribute dynamics. To enhance discrimination, this representation was extended into a Fisher-like encoding that characterizes the first moments of local behavior in the BDS manifold. This generalizes the popular VLAD representation and was denoted the VLADAD. Experiments have shown that the BDS, the BoWAD, and the VLADAD have state of the art performance for activity recognition in video whose segments range from precisely segmented and well aligned to unsegmented and scattered within larger video streams. The ability of these representations to capture signature events of different activity classes was demonstrated through various recounting examples.

Acknowledgements The authors acknowledge valuable discussions on the manuscript and references to variational inference by professor Lawrence K. Saul. This work was partially supported by National Science Foundation (NSF) grant IIS-1208522.

Appendices

A Convergence of Bag-of-Models Clustering

The bag-of-models clustering procedure of Algorithm 2 is a general framework for clustering examples in a Riemannian manifold $\mathcal{M}$ of statistical models. The goal is to find a preset number of models $\{M_j\}_{j=1}^K \subset \mathcal{M}$ in the manifold that best explain a corpora $\mathcal{D} = \{\boldsymbol{z}_i\}_{i=1}^N$ ($\boldsymbol{z}_i \in \mathcal{Z}, \forall i$). It is assumed that all models M are parametrized by a set of parameters $\boldsymbol{\theta}$ and have smooth likelihood functions (derivatives of all orders exist and are bounded), and that Algorithm 2 satisfies the following conditions.

Condition 1: the operation $f_{\mathcal{M}}$ of (20) consists of estimating the parameters $\boldsymbol{\theta}$ of $\mathcal{M}$ by the *maximum likelihood estimation* (MLE) principle.

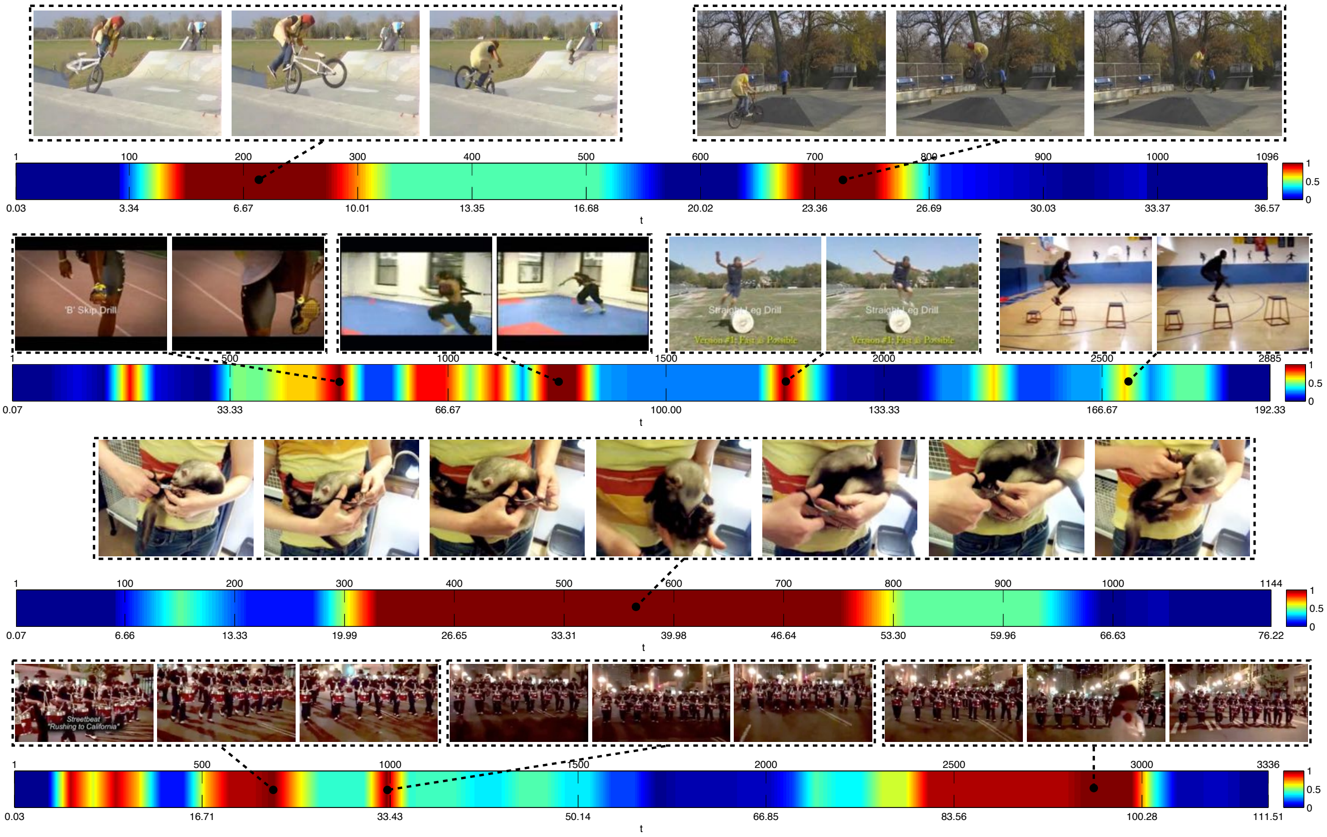


Fig. 16: Recounting by BoWAD on MED11 false positives (top 0.1% detections) for "attempt a board trick," "parkour," "groom an animal," and "parade" (top to bottom).

Condition 2: the Riemannian metric of the manifold $\mathcal{M}$ defined by the Fisher information $\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}}$ (Jaakkola and Haussler, 1999; Amari and Nagaoka, 2000) is used as the dissimilarity measure of (21). More precisely, the metric of $\mathcal{M}$ in the neighbrhood of model $M_{\boldsymbol{z}}$ is

$$d_{\mathcal{M}}(M^*,\ M_{\boldsymbol{z}}) = ||\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}}}, \tag{A.1}$$

where $||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2_{\mathcal{I}} = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\intercal \mathcal{I}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)$, and the Fisher information $\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}}$ is defined as (Amari, 1998)

$$\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}} = -\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x};\boldsymbol{\theta}_{\boldsymbol{z}})}\left[\nabla^2_{\boldsymbol{\theta}} \ln p(\boldsymbol{x};\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{z}}}\right]. \tag{A.2}$$

Given the similarity between Algorithm 2 and k-means, the convergence of the former can be studied with the techniques commonly used to show that the latter converges. This requires the definition of a suitable objective function to quantify the quality of the fit of the set $\{M_i\}_{j=1}^K$ to the corpora $\mathcal{D}$. We rely on the objective

$$\zeta(\{M_i\}_{j=1}^K, \{S_j\}_{j=1}^K) = \sum_j \sum_{\boldsymbol{z}\in S_j} \ln p_{M_j}(\boldsymbol{z}), \tag{A.3}$$

where $p_M(\cdot)$ is the likelihood function of model M, and S_j a subset of $\mathcal{D}$, containing all examples assigned to j-th model. Note that this implies that $\forall i \neq j, S_i \bigcap S_j = \varnothing$ and $\bigcup_j S_j = \mathcal{D}$. From the assumption of smooth models M (*i.e.*, $\forall \boldsymbol{z} \in \mathcal{Z}, M \in \mathcal{M}$, $p_M(\boldsymbol{z}) < \infty$) and the fact that there is only a finite set of assignments $\{S_j\}_{j=1}^K$, the objective function of (A.3) is upper bounded. Since the refinement step of Algorithm 2 updates the models so that

$$M_j^{(t+1)} = f_{\mathcal{M}}(S_j^{(t+1)}) = \arg\max_{M\in\mathcal{M}} \sum_{\boldsymbol{z}\in S_j^{(t+1)}} \ln p_M(\boldsymbol{z}),$$

the objective either increases or remains constant after each refinement step. It remains to prove that the same holds for each assignment step. If that is the case, Algorithm 2 produces a monotonically increasing and upper-bounded sequence of objective function values. By the monotone convergence theorem, this implies that algorithm converges in a finite number of steps. Note that, as in k-means, there is no guarantee on convergence to the global optimum.

It thus remains to prove that the objective of (A.3) increases with each assignment step. The Riemannian structure of the manifold $\mathcal{M}$, makes this proof more technical than the corresponding one for k-means. In what follows, we provide a sketch of the proof. Let M^*

be the model (of parameters $\boldsymbol{\theta}^*$) to which example $\boldsymbol{z}$ is assigned by the assignment step of Algorithm 2, *i.e.*,

$$M^* = \operatorname*{arg\,min}_{M \in \{M_j^{(t)}\}_{j=1}^K} d_{\mathcal{M}}(M_{\boldsymbol{z}}, M) \tag{A.4}$$

and M° (of parameter $\boldsymbol{\theta}^\circ$) the equivalent model of the previous iteration. It follows from Condition 2 that

$$\begin{aligned} d_{\mathcal{M}}(M^*,\ M_{\boldsymbol{z}}) &= ||\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}}} \\ &\leqslant d_{\mathcal{M}}(M^\circ,\ M_{\boldsymbol{z}}) = ||\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}}}. \end{aligned} \tag{A.5}$$

Note that, $M_{\boldsymbol{z}}$ is the model $p(\boldsymbol{z}; \boldsymbol{\theta}_{\boldsymbol{z}})$ onto which $\boldsymbol{z}$ is mapped by (20). From Condition 1, $\boldsymbol{\theta}_{\boldsymbol{z}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{z}; \boldsymbol{\theta})$ and, using a Taylor series expansion,

$$\begin{aligned} \ln p(\boldsymbol{z}; \boldsymbol{\theta}) &\approx \ln p(\boldsymbol{z}; \boldsymbol{\theta}_{\boldsymbol{z}}) + \langle \nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{z}; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{z}}}, \boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{z}} \rangle \\ &\quad + \frac{1}{2} ||\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{H_{\boldsymbol{\theta}_{\boldsymbol{z}}}} && \text{(A.6)} \\ &= \ln p(\boldsymbol{z}; \boldsymbol{\theta}_{\boldsymbol{z}}) + \frac{1}{2} ||\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{H_{\boldsymbol{\theta}_{\boldsymbol{z}}}}, && \text{(A.7)} \end{aligned}$$

where $H_{\boldsymbol{\theta}_{\boldsymbol{z}}} = \nabla^2_{\boldsymbol{\theta}} \ln p(\boldsymbol{z}; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{z}}}$ is the Hessian of $\ln p(\boldsymbol{z}; \boldsymbol{\theta})$ at $\boldsymbol{\theta}_{\boldsymbol{z}}$. Since $p(\boldsymbol{z}; \boldsymbol{\theta}_{\boldsymbol{z}})$ is the model obtained from a single example $\boldsymbol{z}$, it is a heavily peaky distribution centered at $\boldsymbol{z}$. Hence, the expectation of (A.2) can be approximated by

$$\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}} \approx -H_{\boldsymbol{\theta}_{\boldsymbol{z}}}. \tag{A.8}$$

Combining (A.5), (A.7), and (A.8) then results in

$$\begin{aligned} \ln p(\boldsymbol{z}; \boldsymbol{\theta}^*) &\approx \ln p(\boldsymbol{z}; \boldsymbol{\theta}_{\boldsymbol{z}}) + \frac{1}{2} ||\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{H_{\boldsymbol{\theta}_{\boldsymbol{z}}}} \\ &\approx \ln p(\boldsymbol{z}; \boldsymbol{\theta}_{\boldsymbol{z}}) - \frac{1}{2} ||\boldsymbol{\theta}^* - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}}} \\ &\geqslant \ln p(\boldsymbol{z}; \boldsymbol{\theta}_{\boldsymbol{z}}) - \frac{1}{2} ||\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_{\boldsymbol{z}}||^2_{\mathcal{I}_{\boldsymbol{\theta}_{\boldsymbol{z}}}} \approx \ln p(\boldsymbol{z}; \boldsymbol{\theta}^\circ). \end{aligned}$$

It follows that the objective of (A.3) increases after each assignment step. This is intuitive in the sense that, the closer a model M is to an example's representative model, the better M can explain that example.

B Optimization

In this appendix, we derive (72), by considering the optimization problem

$$\begin{aligned} X^* = \operatorname*{arg\,max}_{X \in \mathcal{S}_{++}} & \ b \ln |X| - \operatorname{tr}(AX), \\ & s.t.\ A \in \mathcal{S}_{++},\ \ b > 0. \end{aligned} \tag{B.1}$$

Since 1) both $b \ln |X|$ and $-\operatorname{tr}(AX)$ are smooth and concave functions in X (Boyd and Vandenberghe, 2004), and 2) the domain $\mathcal{S}_{++}$ is an open convex set, the supremum of (B.1) is achieved at either 1) its stationary point(s) (if any), or 2) the boundary of its domain. The derivative of the objective function of (B.1) is

$$\frac{\partial}{\partial X} \{ b \ln |X| - \operatorname{tr}(AX) \} = b (X^{-1})^\intercal - A. \tag{B.2}$$

Setting (B.2) to zero leads to

$$X^* = b A^{-1} \in \mathcal{S}_{++}. \tag{B.3}$$

Applying this result to (71), with $b = 1$, $X = \Sigma$, and $A = W$, leads to (72).

C Variational Inference for BDS

The key computation of the variational inference procedure of Section 6.2.3 is to determine

$$\begin{aligned} \boldsymbol{m}_t &= \langle \boldsymbol{x}_t \rangle_q, \\ \Sigma_{t,t} &= \langle (\boldsymbol{x}_t - \boldsymbol{m}_t)(\boldsymbol{x}_t - \boldsymbol{m}_t)^\intercal \rangle_q, \\ \Sigma_{t,t+1} &= \langle (\boldsymbol{x}_t - \boldsymbol{m}_t)(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{t+1})^\intercal \rangle_q. \end{aligned}$$

In this appendix, we derive an efficient method for this computation, which draws on the solution of the identical variational inference problem for the LDS of (10). We start by discussing the LDS case.

C.1 Inference for Linear Dynamic Systems

Consider the LDS of (10) with parameters $\boldsymbol{\theta}_{LDS} = \{S, \boldsymbol{\mu}, A, C, Q, R, \boldsymbol{u}\}$, an observation sequence $\{\boldsymbol{y}_1^\tau\}$ ($\boldsymbol{y}_t \in \mathbb{R}^K$), and the variational distribution $q(\boldsymbol{x})$ of (57). Similarly to the derivation of Section 6.2.3, the variational lower bound of (41) for the log-likelihood of the LDS can be shown to be

$$\begin{aligned} \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y}, q) &= \langle \ln p(\boldsymbol{x}_1) \rangle_q + \sum_{t=1}^{\tau-1} \langle \ln p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) \rangle_q \\ &\quad + \sum_{t=1}^{\tau} \langle \ln p(\boldsymbol{y}_t|\boldsymbol{x}_t) \rangle_q + H_q(X), \end{aligned} \tag{C.1}$$

with $\langle \ln p(\boldsymbol{x}_1) \rangle_q$, $\langle \ln p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) \rangle_q$, and $H_q(X)$ as in (63)-(65). Furthermore, defining $\tilde{\boldsymbol{y}}_t = \boldsymbol{y}_t - \boldsymbol{u}$,

$$\begin{aligned} \langle \ln p(\boldsymbol{y}_t|\boldsymbol{x}_t) \rangle_q &= \langle \ln \mathcal{G}(\tilde{\boldsymbol{y}}_t; C\boldsymbol{x}_t, R) \rangle_{q(\boldsymbol{x}_t)} \\ &= \langle \ln \mathcal{G}(C\boldsymbol{x}_t; \tilde{\boldsymbol{y}}_t, R) \rangle_{\mathcal{G}(\boldsymbol{x}_t; \boldsymbol{m}_t, \Sigma_{t,t})} \\ &= \langle \ln \mathcal{G}(\boldsymbol{x}_t; \tilde{\boldsymbol{y}}_t, R) \rangle_{\mathcal{G}(\boldsymbol{x}_t; C\boldsymbol{m}_t, C\Sigma_{t,t}C^\intercal)} \end{aligned} \tag{C.2}$$

and, from (8),

$$\langle \ln p(\boldsymbol{y}_t|\boldsymbol{x}_t) \rangle_q \propto -\frac{1}{2} \Big[||\tilde{\boldsymbol{y}}_t - C\boldsymbol{m}_t||^2_R + \operatorname{tr}(R^{-1} C \Sigma_{t,t} C^\intercal) \Big].$$

It follows that

$$\begin{aligned}\mathscr{L}(\boldsymbol{\theta},\boldsymbol{y},q) \propto -\frac{1}{2}\Big\{&||\boldsymbol{\mu}-\boldsymbol{m}_1||_S^2 + \mathrm{tr}(S^{-1}\varSigma_{1,1})\\ &+ \sum_{t=1}^{\tau-1}\mathrm{tr}(\varGamma^{-1}\varPhi_t) + \sum_{t=1}^{\tau}\mathrm{tr}(R^{-1}C\varSigma_{t,t}C^{\intercal})\\ &+ \sum_{t=1}^{\tau}||\tilde{\boldsymbol{y}}_t - C\boldsymbol{m}_t||_R^2\Big\} + \frac{1}{2}\ln|\varSigma|, \qquad \text{(C.3)}\end{aligned}$$

where $\varGamma$ and $\varPhi_t$ are defined in Section 6.2.3.

As was the the case with (69), the optimization of (C.3) with respect to the variational distribution q can be factorized into two optimization problems

$$\begin{aligned}\{\boldsymbol{m}^*,\varSigma^*\} &= \mathop{\arg\max}_{\{\boldsymbol{m},\varSigma\}\in\mathbb{R}^{L\tau}\times\mathcal{S}_{++}^{L\tau}} \mathscr{L}(\boldsymbol{\theta},\boldsymbol{y},q)\\ &= \Big\{\mathop{\arg\max}_{\boldsymbol{m}\in\mathbb{R}^{L\tau}}\mathscr{L}(\boldsymbol{\theta},\boldsymbol{y},q), \mathop{\arg\max}_{\varSigma\in\mathcal{S}_{++}^{L\tau}}\mathscr{L}(\boldsymbol{\theta},\boldsymbol{y},q)\Big\}.\end{aligned}$$

In fact, the dependence of (C.3) on $\varSigma$ is identical to that of (69), up to the replacement of R by $4I$. Hence, the optimal $\varSigma$ is still the solution of (71), *i.e.*,

$$\varSigma^* = W^{-1}, \qquad \text{(C.4)}$$

but with a matrix $W \in \mathcal{S}_{++}$ which is slightly different from (71), namely

$$W_{i,j} = \begin{cases} A^{\intercal}Q^{-1}A + S^{-1} + C^{\intercal}R^{-1}C, & i=j=1,\\ A^{\intercal}Q^{-1}A + Q^{-1} + C^{\intercal}R^{-1}C, & 1<i=j<\tau,\\ Q^{-1} + C^{\intercal}R^{-1}C, & i=j=\tau,\\ -Q^{-1}A, & i=j+1,\\ -A^{\intercal}Q^{-1}, & i=j-1,\\ 0, & \text{otherwise,}\end{cases} \qquad \text{(C.5)}$$

where $W_{i,j} \in \mathbb{R}^{L\times L}$ is the block in row-i, column-j of W.

In summary, the algorithm for learning the covariance of the variational distribution of the BDS is identical to the learning algorithm for the covariance of the variational distribution of a LDS with $R = 4I$. Furthermore, since all random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ (as well as all marginal or conditional distributions) of the LDS are Gaussian, the variational inference is *exact* in this case and

$$q^*(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}_{LDS}).$$

It thus follows that the *standard algorithms* for exact inference of $p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}_{LDS})$ with the LDS can be used to compute the covariance $\varSigma^*$ of the *variational distribution of the BDS*. In the following section, we briefly review the Kalman smoothing filter, which is the most popular such algorithm.

The situation is, however, different for the mean of the variational distribution. In this case, the LDS continues to have a simple closed-form solution, namely

$$\boldsymbol{m}^* = W^{-1}\boldsymbol{\nu}, \qquad \text{(C.6)}$$

where W is as in (C.5) and

$$\boldsymbol{\nu} = \begin{bmatrix}\boldsymbol{\nu}_1\\ \vdots\\ \boldsymbol{\nu}_\tau\end{bmatrix}, \quad \boldsymbol{\nu}_t = \begin{cases} S^{-1}\boldsymbol{\mu} + CR^{-1}\tilde{\boldsymbol{y}}_1, & t=1,\\ CR^{-1}\tilde{\boldsymbol{y}}_t, & 1<t\leqslant\tau.\end{cases} \qquad \text{(C.7)}$$

However, because the dependence of (C.3) on $\boldsymbol{m}$ is no longer identical to that of (69), the LDS solution is not informative for learning the BDS. A different procedure is thus required to learn the variational mean of the BDS. This is discussed in Section C.3.

C.2 Inferring the variational covariance $\varSigma$ of the BDS

Note that, $\boldsymbol{m}^*$ and $\varSigma^*$ have size linear and quadratic, respectively, in τ, the length of the sequence $\{\boldsymbol{y}_1^\tau\}$. This makes the direct solution of (C.6) and (C.4) expensive for long sequences - complexity $O(L^\rho\tau^\rho)$ with $\rho \approx 2.4$. Furthermore, this solution is unnecessary, since inference with both the LDS and BDS only requires $\varSigma_{t,t}^*$ and $\varSigma_{t,t+1}^*$. A popular efficient alternative is the Kalman smoothing filter (Shumway and Stoffer, 1982; Roweis and Ghahramani, 1999), which is commonly use to estimate the posteriors $p(\boldsymbol{x}_t|\boldsymbol{y}_1^\tau)$ and $p(\boldsymbol{x}_t,\boldsymbol{x}_{t+1}|\boldsymbol{y}_1^\tau)$ of the LDS, *i.e.*, $\boldsymbol{m}^*$ of (C.6), $\varSigma_{t,t}^*$ and $\varSigma_{t,t+1}^*$ of (C.4).

Defining expectations conditioned on the observed sequence from time $t=1$ to $t=r$ as

$$\hat{\boldsymbol{x}}_t^r = \langle \boldsymbol{x}_t\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_r)}, \qquad \text{(C.8)}$$

$$\hat{V}_{t,k}^r = \langle(\boldsymbol{x}_t-\hat{\boldsymbol{x}}_t^r)(\boldsymbol{x}_k-\hat{\boldsymbol{x}}_k^r)^{\intercal}\rangle_{p(\boldsymbol{x}_t,\boldsymbol{x}_k|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_r)}, \qquad \text{(C.9)}$$

the estimates are calculated via the *forward* and *backward* recursions:

- In the **forward recursion**, for $t = 1,\cdots,\tau$, compute

$$\hat{V}_{t,t}^{t-1} = A\hat{V}_{t-1,t-1}^{t-1}A^{\intercal} + Q, \qquad \text{(C.10)}$$

$$K_t = \hat{V}_t^{t-1}C^{\intercal}(C\hat{V}_{t,t}^{t-1}C^{\intercal} + R_t)^{-1}, \qquad \text{(C.11)}$$

$$\hat{V}_{t,t}^{t} = \hat{V}_{t,t}^{t-1} - K_tC\hat{V}_{t,t}^{t-1}, \qquad \text{(C.12)}$$

$$\hat{\boldsymbol{x}}_t^{t-1} = A\hat{\boldsymbol{x}}_{t-1}^{t-1}, \qquad \text{(C.13)}$$

$$\hat{\boldsymbol{x}}_t^{t} = \hat{\boldsymbol{x}}_t^{t-1} + K_t(\tilde{\boldsymbol{y}}_t - C\hat{\boldsymbol{x}}_t^{t-1}), \qquad \text{(C.14)}$$

with initial conditions $\hat{\boldsymbol{x}}_1^0 = \boldsymbol{\mu}$ and $\hat{V}_{1,1}^0 = S$.

- In the **backward recursion**, for $t = \tau, \cdots, 1$,

$$J_{t-1} = \hat{V}_{t-1,t-1}^{t-1} A^{\mathsf{T}} (\hat{V}_{t,t}^{t-1})^{-1}, \tag{C.15}$$

$$\hat{\boldsymbol{x}}_{t-1}^{\tau} = \hat{\boldsymbol{x}}_{t-1}^{t-1} + J_{t-1}(\hat{\boldsymbol{x}}_{t}^{\tau} - A\hat{\boldsymbol{x}}_{t-1}^{t-1}), \tag{C.16}$$

$$\hat{V}_{t-1,t-1}^{\tau} = \hat{V}_{t-1,t-1}^{t-1} + J_{t-1}(\hat{V}_{t,t}^{\tau} - \hat{V}_{t,t}^{t-1}) J_{t-1}^{\mathsf{T}}, \tag{C.17}$$

and for $t = \tau, \cdots, 2$,

$$\begin{aligned} \hat{V}_{t-1,t-2}^{\tau} = {} & \hat{V}_{t-1,t-1}^{t-1} J_{t-2}^{\mathsf{T}} \\ & + J_{t-1}(\hat{V}_{t,t-1}^{\tau} - A\hat{V}_{t-1,t-1}^{t-1}) J_{t-2}^{\mathsf{T}} \end{aligned} \tag{C.18}$$

with initial condition $\hat{V}_{\tau,\tau-1}^{\tau} = (I - K_{\tau} C) A \hat{V}_{\tau-1,\tau-1}^{\tau-1}$.

This algorithm can be used to efficiently compute the variational covariance parameters $\Sigma_{t,t}^{*}$ and $\Sigma_{t,t+1}^{*}$ of the BDS, which are exactly the matrices $\hat{V}_{t,t}^{\tau}$ of (C.17) and $\hat{V}_{t,t-1}^{\tau}$ of (C.18), respectively. This has complexity $O(L^{\rho}\tau)$, with $\rho \approx 2.4$.

C.3 Infering the variational mean $\boldsymbol{m}$ of the BDS

The variational mean $\boldsymbol{m}$ is the solution of

$$\begin{aligned} \boldsymbol{m}^{*} &= \arg\max_{\boldsymbol{m}} \ \hat{\mathscr{L}}(\boldsymbol{\theta}, q) \\ &= \arg\max_{\boldsymbol{m}} \Big\{ \boldsymbol{\mu}_0^{\mathsf{T}} S^{-1} \boldsymbol{m}_1 - \frac{1}{2} \boldsymbol{m}_1^{\mathsf{T}} S^{-1} \boldsymbol{m}_1 \\ &\quad - \frac{1}{2} \sum_{t=1}^{\tau-1} \lambda_t^{\mathsf{T}} \Gamma^{-1} \lambda_t \\ &\quad + \sum_{t,k} \Big[\pi_{kt} \ln \sigma(\hat{\omega}_{kt}) + (1 - \pi_{kt}) \ln \sigma(-\hat{\omega}_{kt}) \Big] \Big\}. \end{aligned} \tag{C.19}$$

This can be rewritten as

$$\begin{aligned} \boldsymbol{m}^{*} = \arg\max_{\boldsymbol{m}} \Big\{ & -\frac{1}{2} \boldsymbol{m}^{\mathsf{T}} \tilde{W} \boldsymbol{m} + \boldsymbol{b}_1^{\mathsf{T}} \boldsymbol{m}_1 \\ & + \sum_{t,k} \Big[\pi_{kt} \ln \sigma(\hat{\omega}_{kt}) + (1 - \pi_{kt}) \ln \sigma(-\hat{\omega}_{kt}) \Big] \Big\}, \end{aligned} \tag{C.20}$$

where

$$\tilde{W}_{i,j} = \begin{cases} A^{\mathsf{T}} Q^{-1} A + S^{-1}, & i = j = 1, \\ A^{\mathsf{T}} Q^{-1} A + Q^{-1}, & 1 < i = j < \tau, \\ Q^{-1}, & i = j = \tau, \\ -Q^{-1} A, & i = j + 1, \\ -A^{\mathsf{T}} Q^{-1}, & i = j - 1, \\ 0, & \text{otherwise}, \end{cases} \tag{C.21}$$

$\hat{\omega}_{kt} = C_{k\cdot} \boldsymbol{m}_t + u_k$, and $\boldsymbol{b}_1 = 2S^{-1}\boldsymbol{\mu}$. Since $\hat{\mathscr{L}}(\boldsymbol{\theta}, q)$ is a concave function of $\boldsymbol{m} \in \mathbb{R}^{\tau L}$, gradient-based methods can be applied to search for the stationary point where global optimum is guaranteed.

The gradient of $\hat{\mathscr{L}}(\boldsymbol{\theta}, q)$ is

$$\frac{\partial}{\partial \boldsymbol{m}} \hat{\mathscr{L}}(\boldsymbol{\theta}, q) = -\tilde{W}\boldsymbol{m} + \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} C^{\mathsf{T}} & & \\ & \ddots & \\ & & C^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_{\tau} \end{bmatrix}, \tag{C.22}$$

where

$$\boldsymbol{\beta}_t = \begin{bmatrix} \sigma(\hat{\omega}_{1t}) - \pi_{1t} \\ \vdots \\ \sigma(\hat{\omega}_{Kt}) - \pi_{Kt} \end{bmatrix}.$$

The second-order partial derivatives of $\hat{\mathscr{L}}(\boldsymbol{\theta}, q)$ is

$$\frac{\partial^2}{\partial \boldsymbol{m}^2} \hat{\mathscr{L}}(\boldsymbol{\theta}, q) = -\tilde{W} - \begin{bmatrix} C^{\mathsf{T}} \Xi_1 C & & \\ & \ddots & \\ & & C^{\mathsf{T}} \Xi_{\tau} C \end{bmatrix}, \tag{C.23}$$

where

$$\Xi_t = \operatorname{diag}(\sigma(\hat{\omega}_{1t})\sigma(-\hat{\omega}_{1t}), \ \cdots, \ \sigma(\hat{\omega}_{Kt})\sigma(-\hat{\omega}_{Kt})).$$

Given the concavity and smoothness of $\hat{\mathscr{L}}(\boldsymbol{\theta}, q)$, many popular numerical optimization algorithms can be utilized to search for its optimum, *e.g.*, gradient descent, Newton-Raphson method, Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, *etc.*

D The Fisher Vector for BDS

In this section, we present the derivation of the Fisher vector for BDS using the tightest variational lower bound $\hat{\mathscr{L}}(\boldsymbol{\theta}, q^{*})$ of (69). This consists of computing partial derivatives of $\hat{\mathscr{L}}(\boldsymbol{\theta}, q^{*})$ w.r.t. each of the BDS parameters $\boldsymbol{\theta} = \{S^{-1}, \boldsymbol{\mu}, A, Q^{-1}, C, \boldsymbol{u}\}$.

D.1 Derivative w.r.t. S^{-1}

We have

$$\begin{aligned} & \frac{\partial}{\partial S^{-1}} \hat{\mathscr{L}}(\boldsymbol{\theta}, q^{*}) \\ & = \frac{\partial}{\partial S^{-1}} \frac{1}{2} \Big[\ln |S^{-1}| - \operatorname{tr}\Big((\hat{P}_{1,1}^{*} - 2\boldsymbol{m}_1^{*}\boldsymbol{\mu}^{\mathsf{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}) S^{-1} \Big) \Big] \\ & = \frac{1}{2} \Big(S + 2\boldsymbol{\mu}\boldsymbol{m}_1^{*T} - \hat{P}_{1,1}^{*} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}} \Big), \end{aligned} \tag{D.1}$$

where $\hat{P}_{t_1,t_2}^{*}$ is defined in (62). Note that, $S^{-1} \in \mathcal{S}_{++}^{L}$, thus the derivative of (D.1) needs to be projected into the space of symmetric matrices $\mathcal{S}^{L}$. Since an orthonormal basis of $\mathcal{S}^{L}$ is $\{\frac{1}{2}(E_{i,j} + E_{j,i}), 1 \leqslant i \leqslant j \leqslant L\}$, where $E_{i,j} \in \mathbb{R}^{L \times L}$ with the (i,j)-element equal to one and all the rest elements being zero, it can be shown that after the projection, (D.1) becomes

$$\begin{aligned}&\frac{\partial}{\partial S^{-1}}\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)\\&\quad = \frac{1}{2}\Big(S + \boldsymbol{\mu}\boldsymbol{m}_1^{*T} + \boldsymbol{m}_1^*\boldsymbol{\mu}^\intercal - \hat{P}_{1,1}^* - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal\Big).\end{aligned} \tag{D.2}$$

D.2 Derivative w.r.t. $\boldsymbol{\mu}$

We have

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{\mu}}\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)\\&= \frac{\partial}{\partial \boldsymbol{\mu}}\Big[\boldsymbol{\mu}^\intercal S^{-1}\boldsymbol{m}_1^* - \frac{1}{2}\boldsymbol{\mu}^\intercal S^{-1}\boldsymbol{\mu}\Big]\\&= S^{-1}(\boldsymbol{m}_1^* - \boldsymbol{\mu}).\end{aligned} \tag{D.3}$$

D.3 Derivative w.r.t. A

We have

$$\begin{aligned}&\frac{\partial}{\partial A}\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)\\&= \frac{\partial}{\partial A}\Big[\sum_{t=1}^{\tau-1}\mathrm{tr}\Big(\hat{P}_{t,t+1}^* Q^{-1}A - \frac{1}{2}\hat{P}_{t,t}^* A^\intercal Q^{-1}A\Big)\Big]\\&= \frac{\partial}{\partial A}\Big[\mathrm{tr}\Big(\mathit{\Psi}^\intercal Q^{-1}A - \frac{1}{2}\phi A^\intercal Q^{-1}A\Big)\Big]\\&= (\mathit{\Psi}^\intercal Q^{-1})^\intercal - \frac{1}{2}\Big[Q^{-T}A\phi^\intercal + Q^{-1}A\phi\Big]\\&= Q^{-1}(\mathit{\Psi} - A\phi),\end{aligned} \tag{D.4}$$

where

$$\phi = \sum_{t=2}^{\tau}\hat{P}_{t-1,t-1}^*,\ \mathit{\Psi} = \sum_{t=2}^{\tau}\hat{P}_{t,t-1}^*.$$

D.4 Derivative w.r.t. Q^{-1}

We have

$$\begin{aligned}&\frac{\partial}{\partial Q^{-1}}\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)\\&\quad= \frac{\partial}{\partial Q^{-1}}\Big[\sum_{t=1}^{\tau-1}\mathrm{tr}\Big(A\hat{P}_{t,t+1}^* Q^{-1} - \frac{1}{2}A\hat{P}_{t,t}^* A^\intercal Q^{-1}\\&\qquad\qquad - \frac{1}{2}\hat{P}_{t+1,t+1}^* Q^{-1}\Big) + (\frac{\tau-1}{2})\ln\left|Q^{-1}\right|\Big]\\&\quad= \frac{\partial}{\partial Q^{-1}}\Big[\mathrm{tr}\Big(A\mathit{\Psi}^\intercal Q^{-1} - \frac{1}{2}A\phi A^\intercal Q^{-1} - \frac{1}{2}\varphi Q^{-1}\Big)\\&\qquad\qquad + (\frac{\tau-1}{2})\ln\left|Q^{-1}\right|\Big]\\&\quad= \mathit{\Psi}A^\intercal + \frac{1}{2}\Big[(\tau-1)Q - A\phi A^\intercal - \varphi\Big],\end{aligned} \tag{D.5}$$

where

$$\varphi = \sum_{t=2}^{\tau}\hat{P}_{t,t}^*. \tag{D.6}$$

Again, since $Q^{-1} \in \mathcal{S}_{++}$, the partial derivative of (D.5) is projected into $\mathcal{S}$, giving

$$\begin{aligned}&\frac{\partial}{\partial Q^{-1}}\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)\\&= \frac{1}{2}\Big[\mathit{\Psi}A^\intercal + A\mathit{\Psi}^\intercal - A\phi A^\intercal - \varphi + (\tau-1)Q\Big].\end{aligned} \tag{D.7}$$

D.5 Derivative w.r.t. $\tilde{C}$

Assuming $\tilde{C} = \big[C\ \boldsymbol{u}\big]$, we have

$$\begin{aligned}&\frac{\partial}{\partial \tilde{C}}\hat{\mathscr{L}}(\boldsymbol{\theta}, q^*)\\&= \frac{\partial}{\partial \tilde{C}}\Big\{\sum_{k,t}\Big[\pi_{kt}\ln\sigma(\tilde{C}_{k\cdot}\boldsymbol{b}_t) + (1-\pi_{kt})\ln\sigma(-\tilde{C}_{k\cdot}\boldsymbol{b}_t)\Big]\\&\quad - \frac{1}{8}\mathrm{tr}\big(\tilde{C}\tilde{\mathit{\Upsilon}}\tilde{C}^\intercal\big)\Big\}\\&= -\frac{1}{4}\Big\{\tilde{C}\tilde{\mathit{\Upsilon}} + \sum_{t=1}^{\tau}\begin{bmatrix}\sigma(\tilde{C}_{1\cdot}\boldsymbol{b}_t) - \pi_{1t}\\ \vdots \\ \sigma(\tilde{C}_{K\cdot}\boldsymbol{b}_t) - \pi_{Kt}\end{bmatrix}\boldsymbol{b}_t^\intercal\Big\},\end{aligned} \tag{D.8}$$

Table 7: Examples for Syn-4/5/6

Syn-4	skip-run-walk-wave1
Syn-5	jack-wave1-bend-walk-walk
Syn-6	wave2-run-walk-wave1-jump-wave2

Table 8: Examples for Syn20×1

Ground-truth Activity	wave1-wave1-wave2-walk-walk-wave1-walk-wave2-wave2-walk-jack-skip-wave2-bend-bend-jump-run-skip-jack-wave1
Noisy Instances[1]	side-wave2-walk-skip-run-wave1-bend-bend-walk-walk-wave1-wave1-wave2-walk-walk-wave1-walk-wave2-wave2-walk-jack-skip-wave2-bend-bend-jump-run-skip-jack-wave1-side-bend-side-walk-run-side-walk-jack-bend-walk;
	jump-run-wave1-wave1-wave2-walk-walk-wave1-walk-wave2-wave2-walk-jack-skip-wave2-bend-bend-jump-run-skip-jack-wave1-wave1-walk-side-jump-side-jump-jump-run-jack-side-wave1-run-run-skip-wave1-jack-side-bend;

[1] ground-truth activities are composed of actions in red.

Table 9: Examples for Syn10×2

Ground-truth Activity	jack-jump-side-jump-pjump-run-jack-side-bend-wave1; run-side-side-skip-run-jump-walk-jack-run-skip
Noisy Instances[2]	wave2-run-wave1-bend-jump-wave1-skip-side-jack-jump-side-jump-pjump-run-jack-side-bend-wave1-walk-wave2-wave2-wave1-side-pjump-wave2-run-side-side-skip-run-jump-walk-jack-run-skip-jack-pjump-pjump-pjump-pjump;
	jump-jack-jump-side-jump-pjump-run-jack-side-bend-wave1-jump-side-skip-jack-run-side-bend-jump-pjump-side-run-side-side-skip-run-jump-walk-jack-run-skip-side-pjump-wave2-walk-run-pjump-wave2-wave2-walk;

[2] ground-truth activities are composed of actions in red.

where $\tilde{C}_{k\cdot}$ is the k-th row of $\tilde{C}$, and

$$\tilde{\Upsilon} = \begin{pmatrix} \sum_{t=1}^{\tau} \Sigma^*_{t,t} & 0 \\ 0 & 0 \end{pmatrix}, \ \boldsymbol{b}_t = \begin{pmatrix} \boldsymbol{m}^*_t \\ 1 \end{pmatrix}.$$

E Weizmann Complex Activity

E.1 Synthetic Datasets

The synthetic dataset contains three sets: Syn-4/5/6, Syn20×1 and Syn1s0×2, which are generated using the 10 atomic actions (per person) from the original Weizmann dataset by Gorelick et al (2007). Exemplar activities in Syn-4/5/6, Syn20×1, and Syn10×2 are shown in Table 7, Table 8, and Table 9, respectively. For Syn20×1, and Syn10×2, two of the 9 instances for an activity (each instance is assembled from each of the 9 people's atomic actions).

F Attribute Definition

F.1 Weizmann Complex Activity

Attribute definitions from (Liu et al, 2011) on Weizmann complex activity are shown in Table 10.

F.2 Olympic Sports

Attribute definitions from (Liu et al, 2011) on Olympic Sports dataset (Niebles et al, 2010) are shown in Table 11.

F.3 TRECVID MED11

Attribute definitions from (Bhattacharya, 2013) on TRECVID MED11 dataset (Over et al, 2011) are shown in Table 12.

References

Afsari B, Chaudhry R, Ravichandran A, Vidal R (2012) Group action induced distances for averaging and clustering linear dynamical systems with applications to the analysis of dynamic scenes. CVPR 11

Aggarwal JK, Ryoo MS (2011) Human activity analysis: A review. ACM Computing Surveys 43(16):1–16 1, 3

Amari S, Nagaoka H (2000) Methods of Information Geometry. American Mathematical Society 13, 26

Amari Si (1998) Natural gradient works efficiently in learning. Neural Comput 10(2):251–276 13, 26

Attias H (1999) A variational bayesian framework for graphical models. NIPS 14

Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2010) Action classification in soccer videos with long short-term memory recurrent neural networks. ICANN 4

Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2011) Sequential deep learning for human action recognition. 2nd International Workshop on Human Behavior Understanding 4

Bhattacharya S (2013) Recognition of complex events in open-source web-scale videos: a bottom up approach. ACM International Conference on Multimedia 16, 23, 31

Bhattacharya S, Kalayeh MM, Sukthankar R, Shah M (2014) Recognition of complex events: Exploiting temporal dynamics between underlying concepts. CVPR 5, 19, 23

Blei DM, Lafferty JD (2006) Dynamic topic models. ICML 18

Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Analysis and Machine Intelligence 23(3):257–267 1

Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge University Press 27

Bregler C (1997) Learning and recognizing human dynamics in video sequences. CVPR 1, 3

Campbell L, Bobick A (1995) Recognition of human body motion using phase space constraints. ICCV 3

Chan A, Vasconcelos N (2005) Probabilistic kernels for the classification of auto-regressive visual processes. CVPR 11

Chan A, Vasconcelos N (2008) Modeling, clustering, and segmenting video with mixtures of dynamic textures. IEEE Trans Pattern Analysis and Machine Intelligence 30(5):909–926 12

Chan AB, Vasconcelos N (2007) Classifying video with kernel dynamic textures. CVPR 10

Chang C, Lin C (2011) Libsvm: a library for support vector machines. ACM Trans Intelligent Systems and Technology 2(3):27 16

Chaudhry R, Ravichandran A, Hager G, Vidal R (2009) Histograms of oriented optical flow and binet-cauchy kernels

Table 10: Attributes for Weizmann Actions

attribute	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
arm-hand-alternate-move-forward	0	0	0	0	1	0	0	1	0	0
arm-hand-hang-down-swing-back-forward	0	0	0	0	0	0	0	1	0	0
arm-hand-swing-move-back-forward-motion	0	0	1	0	1	0	1	1	0	0
arm-intense-motion	0	1	1	0	0	0	0	0	0	0
arm-shape-fold	0	0	1	0	1	0	1	0	1	1
arm-shape-straight	1	1	1	1	0	1	0	1	1	1
arm-side-open-up-down-motion	0	1	0	0	0	0	0	0	0	1
arm-small-swing-motion-left-right-up-down	0	1	0	0	0	0	0	0	1	1
arm-synchronized-arm-motion	0	1	1	0	0	0	1	0	0	0
arm-up-motion-over-shoulder	0	1	1	0	0	0	1	0	1	1
chest-level-arm-motion	0	0	0	0	1	0	0	0	0	0
cyclic-motion	0	1	1	1	1	1	1	1	1	1
huge-wave motion-up-down	0	0	1	0	1	1	1	0	0	0
intense-motion	0	1	1	1	1	1	1	0	0	0
leg-alternate-move-forward	0	0	0	0	1	1	0	1	0	0
leg-feet-small-moving-motion	0	0	0	0	0	0	0	1	0	0
leg-intense-motion	0	1	1	1	1	1	1	0	0	0
leg-motion	0	1	1	1	1	1	1	0	0	0
leg-side-stretch-motion	0	1	0	0	0	1	0	0	0	0
leg-two-leg-synchronized-motion	0	1	1	1	0	0	0	0	0	0
leg-up-forward-motion	0	0	1	0	1	0	1	0	0	0
one-arm-motion	1	0	0	0	0	0	1	0	1	0
small-wave-motion-up-down	0	0	0	0	0	0	0	1	0	0
torso-bend-motion	1	0	0	0	0	0	0	0	0	0
torso-vertical-shape-down-forward-motion	0	0	1	0	1	0	1	0	0	0
torso-vertical-shape-down-motion	0	1	0	1	0	0	0	0	0	0
torso-vertical-shape-up-forward-motion	0	0	1	0	1	0	1	0	0	0
torso-vertical-shape-up-motion	0	1	0	1	0	0	0	0	0	0
translation-motion	0	0	1	0	1	1	1	1	0	0
two-arms-motion	0	1	1	0	1	0	0	1	0	1

on nonlinear dynamical systems for the recognition of human actions. CVPR 2, 4, 10, 18

Chomat O, Crowley JL (1999) Probabilistic recognition of activity using local appearance. CVPR 3

Cinbis RG, Verbeek J, Schmid C (2012) Image categorization using fisher kernels of non-iid image models. CVPR 3, 13

Collins M, Dasgupta S, Schapire RE (2002) A generalization of principal component analysis to the exponential family. NIPS 9

Deng L, Yu D (2014) Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing 7(3-4):197–387 4

Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. CVPR 4

Doretto G, Chiuso A, Wu YN, Soatto S (2003) Dynamic textures. International Journal of Computer Vision 51(2):91–109 3, 4, 8, 10

Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: A library for large linear classification. Journal of Machine Learning Research 9:1871–1874 16

Fathi A, Mori G (2008) Action recognition by learning mid-level motion features. CVPR 2

Gaidon A, Harchaoui Z, Schmid C (2011) Actom sequence models for efficient action detection. CVPR 2, 4

Gaidon A, Harchaoui Z, Schmid C (2013) Temporal localization of actions with actoms. IEEE Trans Pattern Analysis and Machine Intelligence 35(11):2782–2795 3

Ghahramani Z, Beal MJ (2000) Propagation algorithms for variational bayesian learning. NIPS 14

Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. IEEE Trans Pattern Analysis and Machine Intelligence 29(12):2247–2253 1, 3, 4, 17, 31

Graves A, Schmidhuber J (2009) Offline handwriting recognition with multidimensional recurrent neural networks. NIPS 4

Graves A, Mohamed Ar, Hinton G (2013) Speech recognition with deep recurrent neural networks. ICASSP 4

Haasdonk B (2005) Feature space interpretation of svms with indefinite kernels. IEEE Trans Pattern Analysis and Machine Intelligence 27(4):482–492 11

Hajimirsadeghi H, Yan W, Vahdat A, Mori G (2015) Visual recognition by counting instances: a multi-instance cardinality potential kernel. CVPR 25

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9(8):1735–1780 4

Ikizler N, Forsyth DA (2008) Searching for complex human activities with no visual examples. International Journal of Computer Vision 80(3):337–357 4

Jaakkola T, Haussler D (1999) Exploiting generative models in discriminative classifiers. NIPS 13, 26

Jain A, Gupta A, Rodriguez M, Davis LS (2013a) Representing videos using mid-level discriminative patches. CVPR 19

Jain M, Jegou H, Bouthemy P (2013b) Better exploiting motion for better action recognition. CVPR 19

Jain M, van Gemert JC, Snoek CGM (2015) What do 15,000 object categories tell us about classifying and localizing actions? CVPR 5

Jegou H, Perronnin F, Douze M, Sanchez J, Perez P, Schmid C (2012) Aggregating local image descriptors into compact codes. IEEE Trans Pattern Analysis and Machine Intelligence 34(9):1704–1718 3, 13, 16

Jhuang H, Gall J, Zuffi S, Schmid C, Black MJ (2013) Towards understanding action recognition. ICCV 5

Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Analysis and Machine Intelligence 35(1):221–231 3, 4

Jiang YG, Dai Q, Xue X, Liu W, Ngo CW (2012) Trajectory-based modeling of human actions with motion reference points. ECCV 3

Jones S, Shao L (2014) A multigraph representation for improved unsupervised/semi-supervised learning of human actions. CVPR 19

Table 11: Attributes for Olympic Sports

attribute	basketball-layup	bowl	clean-jerk	discus-throw	diving-platform-10m	diving-spring-3m	hammer-throw	high-jump	javelin-throw	long-jump	pole-vault	shot-put	snatch	tennis-serve	triple-jump	vault
ball	1	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0
bend	0	1	1	0	0	0	0	0	0	0	0	1	1	1	0	0
big-ball	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
big-step	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
crouch	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0
down-motion-in-air	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0
fast-run	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1
indoor	1	1	1	0	1	1	0	0	0	0	0	0	1	0	0	1
jump	1	0	0	0	1	1	0	1	0	1	1	0	0	0	1	1
jump-forward	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
lift-something	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
local-jump-up	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0
motion-in-the-air	0	0	0	0	1	1	0	1	0	1	1	0	0	0	1	1
one-arm-open	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
one-arm-swing	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
one-hand-holding-pole	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
open-arm-lift	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
outdoor	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1
raise-arms	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
run	1	0	0	0	0	0	0	1	1	1	1	0	0	0	1	1
run-in-air	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
slow-run	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
small-ball	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0
small-local-jump	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
somersault-in-air	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1
spring-platform	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1
standup	0	0	1	0	0	0	0	0	0	1	1	0	1	0	1	0
throw-away	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0
throw-up	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
track	0	1	0	0	0	0	0	0	0	1	1	0	0	0	1	1
turn-around	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0	0
turn-around-with-two-arms-open	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
two-arms-open	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
two-arms-swing-overhead	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
two-hand-holding-pole	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
up-down-motion-local	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
up-motion-in-air	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1
water	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
with-pat	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
with-pole	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0

Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. Machine Learning 37(2):183–233 3, 13

Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. CVPR 4, 16

Kellokumpu V, Zhao G, Pietikäinen M (2008) Human activity recognition using a dynamic texture based method. BMVC 4

Kovashka A, Grauman K (2010) Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. CVPR 4

Krapac J, Verbeek J, Jurie F (2011) Modeling spatial layout with fisher vectors for image categorization. ICCV 3

Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: A large video database for human motion recognition. ICCV 1

Kullback S (1997) Information Theory and Statistics. Courier Dover Publications 8

Lai K, Liu D, Chen M, Chang S (2014) Recognizing complex events in videos by learning key static-dynamic evidences. ECCV 25

Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. CVPR 2, 5

Lan Z, Li X, Hauptmann AG (2014) Temporal extension of scale pyramid and spatial pyramid matching for action recognition. URL http://arxiv.org/abs/1408.7071 3, 20

Lan Z, Lin M, Li X, Hauptmann AG, Raj B (2015) Beyond gaussian pyramid: multi-skip feature stacking for action recognition. CVPR 1, 4, 19, 21

Table 12: Attribute List for TRECVID MED11[1]

Animal-approaching	Animal-eating	Blowdrying	Close-door
Flash-photography	Hands-visible	Machine-carving	Machine-drilling
Machine-planing	Machine-sawing	Open-box	Open-door
People-dancing	People-marching	Perosn-rolling	Person-bending
Person-blowing-candles	Person-carving	Person-casting	Person-cheering
Person-clapping	Person-cleaning	Person-climbing	Person-close-trunk
Person-crying	Person-cutting	Person-cutting-cake	Person-cutting-fabric
Person-dancing	Person-dragging	Person-drilling	Person-drinking
Person-eating	Person-erasing	Person-falling	Person-fitting-bolts
Person-flipping	Person-gluing	Person-hammering	Person-hitting
Person-holding-sword	Person-hugging	Person-inserting-key	Person-jacking-car
Person-jumping	Person-kicking	Person-kissing	Person-laughing
Person-lifting	Person-lighting	Person-lighting-candle	Person-marching
Person-measuring	Person-opening-door	Person-open-trunk	Person-packaging
Person-painting	Person-petting	Person-picking	Person-planing
Person-playing-instrument	Person-pointing	Person-polishing	Person-pouring
Person-pullingout-candles	Person-pushing	Person-pushing	Person-reeling
Person-riding	Person-rolling	Person-running	Person-sawing
Person-sewing	Person-shaking-hands	Person-singing	Person-sketching
Person-sliding	Person-spraying	Person-squatting	Person-standing-up
Person-steering	Person-surfing	Person-taking-pictures	Person-throwing
Person-turning-wrench	Person-twist	Person-twisting-wood	Person-using-knife
Person-using-tire-tube	Person-walking	Person-washing	Person-waving
Person-welding	Person-wetting-wood	Person-whistling	Person-wiping
Person-writing	Shake	Spreading-cream	Stir
Taking-pictures	Vehicle-moving	Wheel-rotating	

[1] About 10,000 short-term clips are annotated for attribute training.

Laptev I (2005) On space-time interest points. International Journal of Computer Vision 64(2-3):107–123 3, 4, 16

Laptev I, Marszałek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. CVPR 1, 3, 4, 12, 18, 19, 23

Laxton B, Lim J, Kriegman D (2007) Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. CVPR 2, 4

Li B, Ayazoglu M, Mao T, Camps O, Sznaier M (2011) Activity recognition using dynamic subspace angles. CVPR 2, 4

Li WX, Vasconcelos N (2012) Recognizing activities by attribute dynamics. NIPS 3, 5

Li WX, Yu Q, Divakaran A, Vasconcelos N (2013a) Dynamic pooling for complex event recognition. ICCV 19

Li WX, Yu Q, Sawhney H, Vasconcelos N (2013b) Recognizing activities via bag of words for attribute dynamics. CVPR 3

Liu J, Kuipers B, Savarese S (2011) Recognizing human actions by attributes. CVPR 2, 3, 5, 16, 17, 19, 31

Matikainen P, Hebert M, Sukthankar R (2010) Representing pairwise spatial and temporal relations for action recognition. ECCV 3

Moore DJ, Essa IA, III MHH (1999) Exploiting human actions and object context for recognition tasks. ICCV 3

Moreno PJ, Ho PP, Vasconcelos N (2004) A kullback-leibler divergence based kernel for svm classification in multimedia applications. NIPS 11

Ng JYH, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. CVPR 3, 4

Ni B, Moulin P, Yang X, Yan S (2015) Motion Part Regularization: Improving Action Recognition via Trajectory Selection. CVPR 1, 4, 19, 21

Niebles JC, Chen CW, Fei-Fei L (2010) Modeling temporal structure of decomposable motion segments for activity classification. ECCV 1, 2, 3, 4, 19, 20, 23, 31

Niyogi S, Adelson E (1994) Analyzing and recognizing walking figures in xyt. CVPR 3

Over P, Awad G, Fiscus J, Antonishek B, Michel M, Smeaton AF, Kraaij W, Quenot G (2011) Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms, and metrics. Proceedings of TRECVID 2011 2, 22, 31

Palatucci M, Pomerleau D, Hinton G, Mitchell T (2009) Zero-shot learning with semantic output codes. NIPS 5

Peng X, Zou C, Qiao Y, Peng Q (2014) Action recognition with stacked fisher vectors. ECCV 1, 4, 16

Peng X, Wang L, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding 150:109–125 4

Perronnin F, Sanchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification. ECCV 3

Pinhanez C, Bobick A (1998) Human action detection using pnf propagation of temporal constraints. CVPR 3

Quattoni A, Collins M, Darrell T (2007) Learning visual representations using images with captions. CVPR 5

Rasiwasi N, Moreno PJ, Vasconcelos N (2007) Bridging the gap: query by semantic example. IEEE Trans Multimedia 9(5):923–938 5

Rasiwasia N, Vasconcelos N (2008) Scene classification with low-dimensional semantic spaces and weak supervision. CVPR 5

Rasiwasia N, Vasconcelos N (2009) Holistic context modeling using semantic co-occurrences. CVPR 2, 5

Rasiwasia N, Vasconcelos N (2012) Holistic context models for visual recognition. IEEE Trans Pattern Analysis and Machine Intelligence 34(5):902–917 2, 4

Ravichandran A, Chaudhry R, Vidal R (2012) Categorizing dynamic textures using a bag of dynamical systems. IEEE Trans Pattern Analysis and Machine Intelligence 35(2):342–353 11, 19

Rodriguez M, Ahmed J, Shah M (2008) Action mach: a spatio-temporal maximum average correlation height filter for action recognition. CVPR 1, 3

Roweis S, Ghahramani Z (1999) A unifying review of linear gaussian models. Neural Computation 11(2):305–345 6, 15, 28

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. URL http://arxiv.org/abs/1409.0575 4

Saul LK, Jordan MI (2000) Attractor dynamics in feedforward neural networks. Neural Computation 12:1313–1335 15

Schein AI, Saul LK, Ungar LH (2003) A generalized linear model for principal component analysis of binary data. AISTATS 3, 9, 10

Schölkopf B, Smola A, Müller K (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10(5):1299–1319 10

Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: A local svm approach. ICPR 1, 3

Shao L, Zhen X, Tao D, Li X (2014) Spatio-temporal Laplacian pyramid coding for action recognition. IEEE Trans Cybernetics 44(6):817–827 3

Shao L, Liu L, Yu M (2015) Kernelized multiview projection for robust action recognition. International Journal of Computer Vision 4

Shumway RH, Stoffer DS (1982) An approach to time series smoothing and forecasting using the em algorithm. Journal of Time Series Analysis 3(4):253–264 8, 28

Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. NIPS 3, 4, 16

Simonyan K, Vedaldi A, Zisserman A (2013) Deep fisher networks for large-scale image classification. NIPS 3

Snoek CGM, Worring M, Smeulders AWM (2005) Early versus late fusion in semantic video analysis. ACM International Conference on Multimedia 21

Sun C, Nevatia R (2013) Active: Activity concept transitions in video event classification. ICCV 5, 19, 23

Sun C, Nevatia R (2014) Discover: Discovering important segments for classification of video events and recounting. CVPR 5

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. CVPR 4

Tamrakar A, Ali S, Yu Q, Liu J, Javed O, Divakaran A, Cheng H, Sawhney H (2012) Evaluation of low-level features and their combinations for complex event detection in open source videos. CVPR 1, 4, 21

Tang K, Fei-Fei L, Koller D (2012) Learning latent temporal structure for complex event detection. CVPR 5, 19, 20, 23

Todorovic S (2012) Human activities as stochastic kronecker graphs. ECCV 19

Vahdat A, Cannons K, Mori G, Oh S, Kim I (2013) Compositional models for video event detection: a multiple kernel learning latent variable approach. ICCV 25

Vasconcelos N, Ho P, Moreno P (2004) The kullback-leibler kernel as a framework for discriminant and localized representations for visual recognition. ECCV 11

Vedaldi A, Zisserman A (2012) Efficient additive kernels via explicit feature maps. IEEE Trans Pattern Analysis and Machine Intelligence 34(3):480–492 16

Vishwanathan SVN, Smola AJ, Vidal R (2006) Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. International Journal of Computer Vision 73(1):95–119 3, 11

Vrigkas M, Nikou C, Kakadiaris I (2015) A review of human activity recognition methods. Frontiers in Robotics and AI 2:1–28 3

Wang H, Schmid C (2013) Action recognition with improved trajectories. ICCV 1, 3, 4, 16, 19, 21

Wang H, Ullah M, Kläser A, Laptev I, Schmid C (2009) Evaluation of local spatio-temporal features for action recognition. BMVC 1, 4, 12

Wang H, Klaser A, Schmid C, Liu CL (2013) Dense trajectories and motion boundary descriptors for action recognition. IJCV 103(1):60–79 3

Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. CVPR 3, 4, 16

Wang X, McCallum A (2006) Topics over time: a non-markov continuous-time model of topical trends. ACM SIGKDD 18

Winn J, Bishop CM (2005) Variational message passing. Journal of Machine Learning Research 6:661–694 14

Xu Z, Tsang I, Yang Y, Ma Z, Hauptmann A (2014) Event detection using multi-level relevance labels and multiple features. CVPR 25

Yacoob Y, Black MJ (1998) Parameterized modeling and recognition of activities. ICCV 3

Ye G, Liu D, Jhuo Ih, Chang Sf (2012) Robust late fusion with rank minimization. CVPR 21

Yu M, Liu L, Shao L (2015) Structure-preserving binary representations for rgb-d action recognition. IEEE Trans Pattern Analysis and Machine Intelligence 3

Yu Q, Liu J, Cheng H, Divakaran A, Sawhney H (2012) Multimedia event recounting with concept based representation. ACM International Conference on Multimedia 22