Multi-view latent variable discriminative models for action recognition

Citation: Y. Song, L. Morency, and R. Davis. “Multi-View Latent Variable Discriminative Models for Action Recognition.” 2012 IEEE Conference on Computer Vision and Pattern Recognition (n.d.). doi:10.1109/cvpr.2012.6247918.

As Published: http://dx.doi.org/10.1109/CVPR.2012.6247918

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: http://hdl.handle.net/1721.1/86101

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Appeared in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012.*

Multi-View Latent Variable Discriminative Models For Action Recognition

Yale Song[1], Louis-Philippe Morency[2], Randall Davis[1]
[1]MIT Computer Science and Artificial Intelligence Laboratory
[2]USC Institute for Creative Technology
{yalesong,davis}@csail.mit.edu, morency@ict.usc.edu

Abstract

Many human action recognition tasks involve data that can be factorized into multiple views such as body postures and hand shapes. These views often interact with each other over time, providing important cues to understanding the action. We present multi-view latent variable discriminative models that jointly learn both view-shared and view-specific sub-structures to capture the interaction between views. Knowledge about the underlying structure of the data is formulated as a multi-chain structured latent conditional model, explicitly learning the interaction between multiple views using disjoint sets of hidden variables in a discriminative manner. The chains are tied using a predetermined topology that repeats over time. We present three topologies – linked, coupled, and linked-coupled – that differ in the type of interaction between views that they model. We evaluate our approach on both segmented and unsegmented human action recognition tasks, using the ArmGesture, the NATOPS, and the ArmGesture-Continuous data. Experimental results show that our approach outperforms previous state-of-the-art action recognition models.

1. Introduction

Many real-world human action recognition tasks involve data that can be factorized into multiple views. For example, the gestures made by baseball coaches involve complex combinations of body and hand signals. The use of multiple views in human action recognition has been shown to improve recognition accuracy [1, 3]. Evidence from psychological experiments provides theoretical justification [25], showing that people reason about interaction between views (i.e., causal inference) when given combined input signals. We introduce the term *multi-view dynamic learning* as a mechanism for such tasks. The task involves sequential data, where each view is generated by a temporal process and encodes a different source of information. These views often exhibit both view-shared and view-specific substructures [11], and usually interact with each other over time, providing important cues to understanding the data.

Single-view latent variable discriminative models (e.g., HCRF [18] for segmented sequence data, and LDCRF [15] for unsegmented sequence data) have shown promising results in many human activity recognition tasks such as gesture and emotion recognition. However, when applied to multi-view latent dynamic learning, existing latent models (e.g., early fusion [27]) often prove to be inefficient or inappropriate. The main difficulty with this approach is that it needs a set of latent variables that are the product set of the latent variables from each original view [16]. This increase in complexity is exponential: with C views and D latent variables per view, the product set of all latent variables is $O(D^C)$. This in turn causes the model to require much more data to estimate the underlying distributions correctly (as confirmed in our experiment shown in Section 4), which makes this solution impractical for many real world applications. The task can get even more difficult when, as shown in [5], one process with high dynamics (e.g., high variance, noise, frame rate) masks another with low dynamics, with the result that both the view-shared and view-specific substructures are dominated by the view with high dynamics.

We present here multi-view latent variable discriminative models that jointly learn both view-shared and view-specific sub-structures. Our approach makes the assumption that observed features from different views are conditionally independent given their respective sets of latent variables, and uses disjoint sets of latent variables to capture the interaction between views. We introduce multi-view HCRF (MV-HCRF) and multi-view LDCRF (MV-LDCRF) models, which extend previous work on HCRF [18] and LDCRF [15] to the multi-view domain (see Figure 1).

Knowledge about the underlying structure of the data is represented as a multi-chain structured conditional latent model. The chains are tied using a predetermined topology that repeats over time. Specifically, we present three topologies –linked, coupled, and linked-coupled– that differ in the type of interaction between views that they model. We demonstrate the superiority of our approach over existing single-view models using three real world human action datasets – the ArmGesture [18], the NATOPS [22], and the

ArmGesture-Continuous datasets – for both segmented and unsegmented human action recognition tasks.

Section 2 reviews related work, Section 3 presents our models, Section 4 demonstrates our approach using synthetic example, and Section 5 describes experiments and results on the real world data. Section 6 concludes with our contributions and suggests directions for future work.

2. Related Work

Conventional approaches to multi-view learning include early fusion [27], i.e., combining the views at the input feature level, and late fusion [27], i.e., combining the views at the output level. But these approaches often fail to learn important sub-structures in the data, because they do not take multi-view characteristics into consideration.

Several approaches have been proposed to exploit the multi-view nature of the data. Co-training [2] and multiple kernel learning [14, 23] have shown promising results when the views are independent, i.e., they provide different and complementary information of the data. However, when the views are not independent, as is common in human activity recognition, these methods often fail to learn from the data correctly [12]. Canonical correlation analysis (CCA) [8] and sparse coding methods [11] have shown a powerful generalization ability to model dependencies between views. However, these approaches are applicable only to classification and regression problems, and cannot be applied directly to dynamic learning problems.

Probabilistic graphical models have shown to be extremely successful in dynamic learning. In particular, multi-view latent dynamic learning using a generative model (e.g., HMM) has long been an active research area [3, 16]. Brand *et al.* [3] introduced a coupled HMM for action recognition, and Murphy introduced Dynamic Bayesian Networks [16] that provide a general framework for modeling complex dependencies in hidden (and observed) state variables.

In a discriminative setting, Sutton *et al.* [24] introduced a dynamic CRF (DCRF), and presented a factorial CRF as an instance of the DCRF, which performs multi-labeling tasks. However, their approach only works with single-view input, and may not capture the sub-structures in the data because it does not use latent variables [18]. More recently, Chen *et al.* presented a multi-view latent space Markov Network for multi-view object classification and annotation tasks [4].

Our work is different from the previous work in that, instead of making the view independence assumption as in [2, 14, 23], we make a conditional independence assumption between views, maintaining computational efficiency while capturing the interaction between views. Unlike [24], we formulate our graph to handle multiple input streams independently, and use latent variables to model the substructure of the multi-view data. We also concentrate on multi-view dynamic sequence modeling, as compared to multi-view object recognition [4, 8, 11, 14, 23].

3. Our Multi-view Models

In this section we describe our multi-view latent variable discriminative models. In particular, we introduce two new family of models, called multi-view HCRF (MV-HCRF) and multi-view LDCRF (MV-LDCRF) that extend previous work on HCRF [18] and LDCRF [15] to the multi-view domain. The main difference between the two models lies in that the MV-HCRF is for segmented sequence labeling (i.e., one label per sequence) while the MV-LDCRF is for unsegmented sequence labeling (i.e., one label per frame). We first introduce the notation, describe MV-HCRF and MV-LDCRF, present three topologies that define how the views interact, and explain inference and parameter estimation.

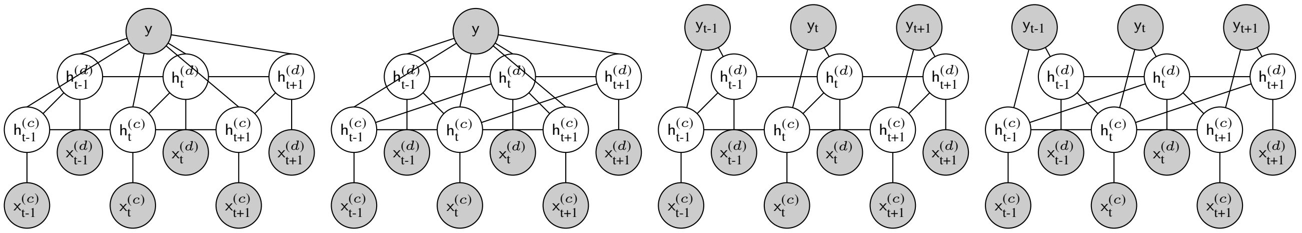
Input to our model is a set of multi-view sequences $\hat{\mathbf{x}} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(C)}\}$, where each $\mathbf{x}^{(c)} = \{\mathbf{x}_1^{(c)}, \cdots, \mathbf{x}_T^{(c)}\}$ is an observation sequence of length T from the c-th view. Each $\hat{\mathbf{x}}_t$ is associated with a label y_t that is a member of a finite discrete set $\mathcal{Y}$; for segmented sequences, there is only one y for all t. We represent each observation $\mathbf{x}_t^{(c)}$ with a feature vector $\phi(\mathbf{x}_t^{(c)}) \in \mathbb{R}^N$. To model the sub-structure of the multi-view sequences, we use a set of latent variables $\hat{\mathbf{h}} = \{\mathbf{h}^{(1)}, \cdots, \mathbf{h}^{(C)}\}$, where each $\mathbf{h}^{(c)} = \{h_1^{(c)}, \cdots, h_T^{(c)}\}$ is a hidden state sequence of length T. Each random variable $h_t^{(c)}$ is a member of a finite discrete set $\mathcal{H}^{(c)}$ of the c-th view, which is disjoint from view to view. Each hidden variable $h_s^{(c)}$ is indexed by a pair (s, c). An edge between two hidden variables $h_s^{(c)}$ and $h_t^{(d)}$ is indexed by a quadruple (s, t, c, d), where $\{s, t\}$ describes the time indices and $\{c, d\}$ describes the view indices.

3.1. Multi-view HCRF

We represent our model as a conditional probability distribution that factorizes according to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_P, \mathcal{E}_S)$ defined over a multi-chain structured stochastic process, where each chain is a discrete representation of each view. A set of vertices $\mathcal{V}$ represents random variables (observed or unobserved) and the two sets of edges $\mathcal{E}_P$ and $\mathcal{E}_S$ represent dependencies among random variables. The unobserved (hidden) variables are marginalized out to compute the conditional probability distribution. We call $\mathcal{E}_P$ a set of *view-specific* edges; they encode temporal dependencies specific to each view. $\mathcal{E}_S$ is a set of *view-shared* edges that encode interactions between views.

Similar to HCRF [18], we construct a conditional probability distribution with a set of weight parameters $\Lambda = \{\lambda, \omega\}$ as

$$p(y \mid \hat{\mathbf{x}}; \Lambda) = \sum_{\hat{\mathbf{h}}} p(y, \hat{\mathbf{h}} \mid \hat{\mathbf{x}}; \Lambda) = \frac{1}{Z} \sum_{\hat{\mathbf{h}}} e^{\Lambda^\intercal \Phi(y, \hat{\mathbf{h}}, \hat{\mathbf{x}})} \tag{1}$$



(a) Linked HCRF (b) Coupled HCRF (c) Linked LDCRF (d) Coupled LDCRF

Figure 1. **Graphical representations of multi-view latent variable discriminative models:** (a) linked HCRF (LHCRF), (b) coupled HCRF (CHCRF), (c) linked LDCRF (LLDCRF), and (d) coupled LDCRF (CLDCRF). Grey nodes are observed variables and white nodes are unobserved variables. The topologies (i.e., linked and coupled) differ by modeling the type of interaction between views; see the text for detail. The linked-coupled multi-view topologies (not shown here) are a combination of the linked and coupled topologies. Note that we illustrate two-view models for simplicity, generalization to >2 views can be done easily by following the rules stated in Equation 5.

where $\Lambda^{\mathsf{T}}\Phi(y,\hat{\mathbf{h}},\hat{\mathbf{x}})$ is a *potential function* and $Z = \sum_{y'\in\mathcal{Y},\hat{\mathbf{h}}} e^{\Lambda^{\mathsf{T}}\Phi(y',\hat{\mathbf{h}},\hat{\mathbf{x}})}$ is a *partition function* for normalization. The potential function is factorized with feature functions $f_k(\cdot)$ and $g_k(\cdot)$ as

$$\begin{aligned}\Lambda^{\mathsf{T}}\Phi(y,\hat{\mathbf{h}},\hat{\mathbf{x}}) = & \sum_{(s,c)\in\mathcal{V}}\sum_k \lambda_k f_k(y,h_s^{(c)},\mathbf{x}^{(c)}) \\ & + \sum_{(s,t,c,d)\in\mathcal{E}}\sum_k \omega_k g_k(y,h_s^{(c)},h_t^{(d)},\hat{\mathbf{x}})\end{aligned} \tag{2}$$

where $\mathcal{E} = \mathcal{E}_P \cup \mathcal{E}_S$. The first term $\lambda_k f_k(\cdot)$ represents singleton potentials defined over a single hidden variable $h_s^{(c)} \in \mathcal{V}$, and the second term $\omega_k g_k(\cdot)$ represents pairwise potentials over a pair of hidden variables $(h_s^{(c)}, h_t^{(d)}) \in \mathcal{E}$.

We define two types of $f_k(\cdot)$ feature functions. The *label* feature function $f_k(y, h_s^{(c)})$ models the relationship between a hidden state $h_s^{(c)} \in \mathcal{H}^{(c)}$ and a label $y \in \mathcal{Y}$; thus, the number of the label feature functions is $\sum_c |\mathcal{Y}| \times |\mathcal{H}^{(c)}|$. The *observation* feature function $f_k(h_s^{(c)}, \mathbf{x}^{(c)})$ represents the relationship between a hidden state $h_s^{(c)} \in \mathcal{H}^{(c)}$ and observations $\phi(\mathbf{x}^{(c)})$, and is of length $\sum_c |\mathcal{H}^{(c)}| \times |\phi(\mathbf{x}^{(c)})|$. Note that $f_k(\cdot)$ are modeled under the assumption that views are conditionally independent given hidden variables, and thus encode the view-specific sub-structures.

The feature function $g_k(\cdot)$ encodes both view-shared and view-specific sub-structures. The definition of $g_k(\cdot)$ depends on how the two sets of edges $\mathcal{E}_P$ and $\mathcal{E}_S$ are defined; we detail these in Section 3.3.

Once we obtain the optimal set of parameters Λ^* (described in Section 3.4), a class label y^* for a new observation sequence $\hat{\mathbf{x}}$ is determined as $y^* = \arg\max_{y\in\mathcal{Y}} p(y \mid \hat{\mathbf{x}};\ \Lambda^*)$.

Note that multi-view HCRF is similar to HCRF [18]; the difference lies in the multi-chain structured model (c.f., the tree-structured model in HCRF), which makes our model capable of multi-view dynamic learning.

3.2. Multi-view LDCRF

The MV-HCRF models described above have focused on the task of segmented sequence labeling, where only one label y is assigned to the whole sequence. In this section, we propose a second family of multi-view models tailored to unsegmented sequence labeling, called multi-view LDCRF (MV-LDCRF), which is inspired by LDCRF [15].

An MV-LDCRF is a multi-view discriminative model for simultaneous sequence segmentation and labeling that can capture both intrinsic and extrinsic class dynamics. Similar to [15], we assume that each class label $y_t \in \mathcal{Y}$ has a disjoint set of associated hidden states $\hat{\mathcal{H}}_y$, which makes $p(\mathbf{y} \mid \hat{\mathbf{h}}, \hat{\mathbf{x}}; \Lambda) = 0$ for any $h_s^{(c)} \notin \mathcal{H}_{y_s}^{(c)}$. Therefore, a conditional probability distribution is written as:

$$p(\mathbf{y} \mid \hat{\mathbf{x}}; \Lambda) = \sum_{\hat{\mathbf{h}}:\forall h_s^{(c)} \in \mathcal{H}_{y_s}^{(c)}} p(\hat{\mathbf{h}} \mid \hat{\mathbf{x}}; \Lambda) \tag{3}$$

The definition of feature functions $f_k(\cdot)$ and $g_k(\cdot)$ are similar to MV-HCRF (see Section 3.1); the only difference is that we include only the observation function for $f_k(\cdot)$, i.e., no label feature function $f_k(y, h_s^{(c)})$.

For testing, instead of estimating a single most probable sequence label y^*, we want to estimate a sequence of most probable labels $\mathbf{y}^*$. This is obtained as:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}\in\mathcal{Y}} \sum_{c=1}^{C} \sum_{\mathbf{h}^{(c)}:\forall h_s^{(c)} \in \mathcal{H}_{y_s}^{(c)}} \alpha_c p(\mathbf{h}^{(c)} \mid \hat{\mathbf{x}}; \Lambda^*) \tag{4}$$

where C is the number of views, and $\sum_c \alpha_c = 1$ sets relative weights on the marginal probability from the c-th view. In our experiments, since we had no prior knowledge about the relative importance of each view, we set all α_c to $1/C$. To estimate the label y_t^* of the t-th frame, we compute the marginal probabilities $p(h_t^{(c)} = h' \mid \hat{\mathbf{x}}; \Lambda^*)$ for all views $c \in C$ and for all hidden states $h' \in \mathcal{H}^{(c)}$ of each view.

Then, for each view c, we sum the marginal probabilities according to $\mathcal{H}_{y_s}^{(c)}$, and compute a weighted mean of them across all views. Finally, the label y_t' that is associated with the optimal set is chosen.

3.3. Topologies: Linked and Coupled

The configuration of $\mathcal{E}_P$ and $\mathcal{E}_S$ encodes the view-shared and view-specific sub-structures. Inspired by [16], we present three topologies that differ in defining the view-shared edges $\mathcal{E}_S$: linked, coupled, and linked-coupled. Because these topologies have repeating patterns, they make the algorithm simple yet powerful, as in HCRF [18]. Figure 1 illustrates graphical representations of linked and coupled topologies for both the MV-HCRF and MV-LDCRF families.

The *linked* multi-view topologies (Figure 1(a) and 1(c)) model contemporaneous connections between views, i.e., the current state in one view concurrently affects the current state in the other view. Intuitively, this captures the synchronization points between views. The *coupled* multi-view topologies (Figure 1(b) and 1(d)) model first-order Markov connections between views, i.e., the current state in one view affects the next state in the other view. Intuitively, this captures the "poker game" interaction, where one player's move is affected by the other player's previous move (but no synchronization between players at the current move). The *linked-coupled* multi-view topologies are a combination of the linked and coupled topologies, i.e., it models the "full" interaction between each pair of multiple sequential chains. In all our models, we assume view-specific first-order Markov chain structure, i.e., the current state affects the next state in the same view.

We encode these dependencies by defining the *transition* feature functions $g_k(\cdot)$ as

$$g_k(y, h_s^{(c)}, h_t^{(d)}) = \mathbf{1} \iff \begin{cases} (s+1=t \wedge c=d) \vee (s=t \wedge c \neq d) & \text{(linked)} \\ (s+1=t) & \text{(coupled)} \\ (s+1=t) \vee (s=t \wedge c \neq d) & \text{(linked-coupled)} \end{cases} \quad (5)$$

In other words, each feature $g_k(\cdot)$ is non-zero only when there is an edge between $h_s^{(c)}$ and $h_t^{(d)}$ as specified in the union of $\mathcal{E}_P$ and $\mathcal{E}_S$.

3.4. Parameter Estimation and Inference

Given a training dataset $\mathcal{D} = \{y_i, \hat{\mathbf{x}}_i\}_{i=1}^N$, we find the optimal parameter set $\Lambda^* = \{\lambda^*, \omega^*\}$ by minimizing the conditional log-likelihood [1]

$$\min_{\Lambda} L(\Lambda) = \frac{\gamma}{2}\|\Lambda\|^2 - \sum_{i=1}^{N} \log p(y_i \mid \hat{\mathbf{x}}_i;\ \Lambda) \qquad (6)$$

where the first term is an L2-norm regularization factor. The second term, inside the summation, can be re-written as:

$$\log p(y_i|\hat{\mathbf{x}}_i;\Lambda) = \sum_{\hat{\mathbf{h}}} \Lambda^\intercal \Phi(y,\hat{\mathbf{h}},\hat{\mathbf{x}}) - \sum_{y',\hat{\mathbf{h}}} \Lambda^\intercal \Phi(y',\hat{\mathbf{h}},\hat{\mathbf{x}}) \quad (7)$$

As with other latent models (e.g., HMMs [19]), introducing hidden variables in Equation 7 makes our objective function non-convex. We find the optimal parameters Λ^* using the recently proposed non-convex regularized bundle method (NRBM) [7], which has been proven to converge to a solution with an accuracy ϵ at the rate $O(1/\epsilon)$. The method aims at iteratively building an increasingly accurate piecewise quadratic lower bound of $L(\Lambda)$ based on its subgradient $\partial_\Lambda L(\Lambda) = \partial_\lambda L(\Lambda) + \partial_\omega L(\Lambda)$. $\partial_\lambda L(\Lambda)$ can be computed as follows ($\partial_\omega L(\Lambda)$ omitted for space):

$$\begin{aligned} \frac{\partial L(\Lambda)}{\partial \lambda_k} &= \sum_{(s,c),h'} p(h_s^{(c)} = h' \mid y, \mathbf{x}^{(c)}; \Lambda) f_k(\cdot) \qquad (8) \\ &- \sum_{(s,c),h',y'} p(y', h_s^{(c)} = h' \mid \mathbf{x}^{(c)}; \Lambda) f_k(\cdot) \end{aligned}$$

The most computationally intensive part of solving Equation 6 is the inference task of computing the marginal probabilities in Equation 8. We implemented both the junction tree (JT) algorithm [6] for an exact inference and the loopy belief propagation (LBP) [17] for an efficient approximate inference. For LBP, the message update is done with random scheduling. The update is considered as "converged" when the previous and the current marginals differ by less than 10^{-4}.

Note that we can easily change our optimization problem in Equation 6 into the max-margin approach [26] by replacing $\sum_{\hat{\mathbf{h}}}$ and $\sum_{y',\hat{\mathbf{h}}}$ in Equation 7 with $\max_{\hat{\mathbf{h}}}$ and $\max_{y',\hat{\mathbf{h}}}$ and solving MAP inference problem. Max-margin approaches have recently been shown to improve the performance of HCRF [26]; we plan to implement the max-margin approach for our multi-view models in the future.

4. Experiments on Synthetic Data

As an initial demonstration of our approach, we use a synthetic example and compare three MV-HCRF models (LHCRF, CHCRF, and LCHCRF) to a single-view HCRF. Specifically, we focus on showing the advantage of exponentially reduced model complexity of our approach by comparing performance with varying training dataset size.

Dataset: Synthetic data for a three-view binary classification task was generated using the Gibb's sampler [21]. Three first-order Markov chains, with 4 hidden states each, were tied using the linked-coupled topology (i.e., LCHCRF; see Equation 5). In order to simulate three views having strong interaction, we set the weights on view-shared edges to random real values in the range $[-10, 10]$, while view-specific edge weights were $[-1, 1]$. The observation model

[1] The derivations in this section is for MV-HCRF. This can be changed easily for MV-LDCRF by replacing y_i with $\mathbf{y}_i$.

for each view was defined by a 4D exponential distribution; to simulate three views having different dynamics, the min-max range of the distribution was set to $[-1, 1]$, $[-5, 5]$, and $[-10, 10]$, respectively. To draw samples, we randomly initialized model parameters, and iterated 50 times using the Gibb's sampler. The sequence length was set to 30.

Methodology: Inference in the MV-HCRF models was performed using both the junction tree and the loopy BP algorithms. Since we know the exact number of hidden states per view, the MV-HCRF models were set to have that many hidden states. The optimal number of hidden states for the HCRF was selected automatically based on validation, varying the number from 12 to 72 with an increment of 12. The size of the training and validation splits were varied from 100 to 500 with an increment of 100, the size of the test split was always 1,000. For each split size, we performed 5-fold cross validation except that the test split size stayed constant at 1,000. Each model was trained with five random initializations, and the best validation parameter was selected based on classification accuracy.

Result: Figure 2 and Table 1 show classification accuracy as a function of dataset size, comparing HCRF and the three MV-HCRF models. The results are averaged values over the 5 splits. The MV-HCRF models always outperformed the HCRF; these differences were statistically significant for the dataset size of 100 and 300. Our result also shows that an approximate inference method (i.e., LBP) on the multi-view models achieves as good accuracy as done with an exact inference method (i.e., JT).

The optimal number of hidden states for the best performing HCRF was 48, which resulted in a sizable difference in the model complexity, with 6,144 parameters to estimate using HCRF, compared to the number of parameters needed for LHCRF (240), CHCRF (336), and LCHCRF (432). Consequently, the MV-HCRF models outperformed the HCRF consistently even with one third of the training dataset size (see Table 1, dataset size 100 and 300). Note that the performance difference between multi-view and single-view HCRFs is larger at a smaller training split size. This implies that our multi-view approach is advantageous especially when there is not enough training data, which is often the case in practice.

5. Experiments on Real-world Data

We evaluated our multi-view models on segmented and unsegmented human action recognition tasks using three datasets: the ArmGesture dataset [18], the NATOPS dataset [22], and the ArmGesture-Continuous dataset.[2] The first two datasets involve segmented gestures, while the third involves unsegmented gestures that we created based on [18]. Below we describe the datasets, detail our experimental methods with baselines, and report and discuss the results.

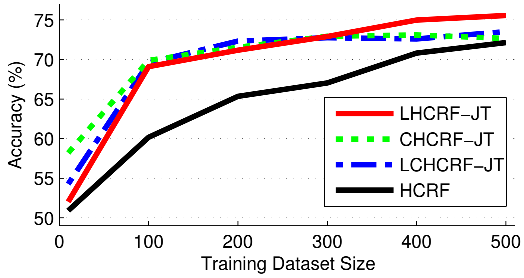


Figure 2. **Accuracy graph as a function of the training dataset size on the synthetic dataset.**

Models	Accuracy (%)		
	N=100	N=300	N=500
LHCRF-JT	**69.58** (p<.01)	**72.76** (p=.01)	73.52 (p=.12)
CHCRF-JT	**69.88** (p<.01)	**72.94** (p<.01)	72.66 (p=.56)
LCHCRF-JT	**69.10** (p<.01)	**72.92** (p<.01)	**75.56** (p=.03)
LHCRF-LBP	**69.94** (p<.01)	**72.70** (p=.01)	**75.28** (p=.01)
CHCRF-LBP	**69.66** (p<.01)	**71.50** (p=.01)	72.60 (p=.59)
LCHCRF-LBP	**68.86** (p<.01)	**72.22** (p=.03)	73.44 (p=.17)
HCRF	60.18	67.04	72.14

Table 1. **Experimental results on the synthetic dataset.** The MV-HCRF models statistically significantly outperformed the single-view HCRF at the dataset size of 100 and 300. Values in parenthesis show p-values from t-tests against HCRF. Bold faced values indicate the difference against HCRF was statistically significant.

5.1. Datasets

ArmGesture [18]: This dataset includes the six arm gestures shown in Figure 3. Observation features include automatically tracked 2D joint angles and 3D euclidean coordinates for left/right shoulders and elbows; each observation is represented as a 20D feature vector. The dataset was collected from 13 participants with an average of 120 samples per class.[3] Following [20], we subsampled the data by the factor of 2. For multi-view models, we divided signals into the left and right arms.

NATOPS [22]: This dataset includes twenty-four body-hand gestures used when handling aircraft on the deck of an aircraft carrier. We used the six gestures shown in Figure 4. The dataset includes automatically tracked 3D body postures and hand shapes. The body feature includes 3D joint velocities for left/right elbows and wrists, and represented as a 12D input feature vector. The hand feature includes probability estimates of five predefined hand shapes – opened/closed palm, thumb up/down, and "no hand". The fifth shape, no hand, was dropped in the final representa-

[2] The dataset and the source code of our model are available at `http://people.csail.mit.edu/yalesong/cvpr12/`

[3] The exact sample counts per class were [88, 117, 118, 132, 179, 90].

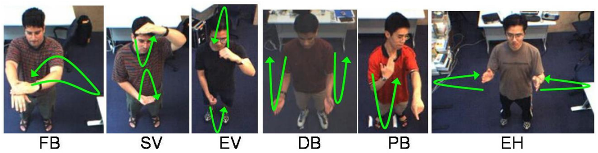


Figure 3. **ArmGesture dataset [18].** Illustration of the 6 gesture classes (Flip Back, Shrink Vertically, Expand Vertically, Double Back, Point and Back, Expand Horizontally). The green arrows are the motion trajectory of the fingertip.

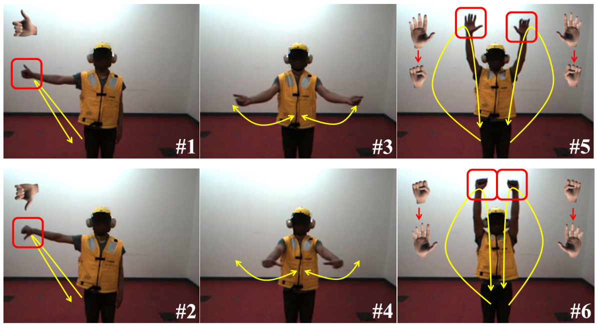


Figure 4. **NATOPS dataset [22].** Illustration of the 6 aircraft handling signal gestures. Body movements are illustrated in yellow arrows, and hand poses are illustrated with synthesized images of hands. Red rectangles indicate hand poses are important in distinguishing the gesture pair.

tion, resulting in an 8D input feature vector (both hands). Note that, in the second gesture pair (#3 and #4), the hand signals contained mostly zero values, because these hand shapes were not included in the predefined hand poses. We included this gesture pair to see how the multi-view models would perform when one view has almost no information. We used 10 samples per person, for a total of 200 samples per class. Similar to the previous experiment, we subsampled the data by the factor of 2. For multi-view models, we divided signals into body postures and hand shapes.[4]

ArmGesture-Continuous: The two above mentioned datasets include segmented sequences only. To evaluate the MV-LDCRF models on unsegmented sequences, we created a new dataset based on the ArmGesture dataset, called ArmGesture-Continuous. To generate an unsegmented sequence, we randomly selected 3 to 5 (segmented) samples from different classes, and concatenated them in random order. This resulted in 182 samples in total, with an average of 92 frames per sample. Similar to the previous experiment, we subsampled the data by the factor of 2, and divided signals into the left and right for multi-view models.

5.2. Methodology

Based on the previous experiment on the synthetic data, we selected two topologies, linked and coupled, to compare to several baselines. Also, the junction tree algorithm is selected as an inference method for the multi-view models. In all experiments, we performed four random initializations of the model parameters, and the best validation parameter was selected based on classification accuracy. Below we detail experimental methods and baselines used in each experiment.

ArmGesture: Five baselines were chosen: HMM [19], CRF [13], HCRF [18], max-margin HCRF [26], and S-KDR-SVM [20]. Following [20], we performed 5-fold cross validation.[5] The number of hidden states was automatically validated; for the single-view model (i.e., MM-HCRF), we varied it from 8 to 16, increasing by 4; and for multi-view models, we varied it from 8 (4 per view) to 16 (8 per view), increasing by 4 (2 per view). No regularization was used in this experiment.

NATOPS: Three baselines were chosen: HMM [19], CRF [13], and HCRF [18]. We performed hold-out testing, where we selected samples from the last 10 subjects for training, the first 5 subjects for testing, and the remaining 5 subjects for validation. The number of hidden states was automatically validated; for single-view models (i.e., HMM and HCRF), we varied it from 12 to 120, increasing by 12; for multi-view models, we varied it from 6 (3 per view) to 60 (30 per view). No regularization was used in this experiment.

ArmGesture-Continuous: Unlike the two previous experiments, the task in this experiment was continuous sequence labeling. We compared a linked LDCRF to two baselines: CRF [13] and LDCRF [15]. We performed hold-out testing, where we selected the second half of the dataset for training, the first quarter for testing, and the remaining for validation. The number of hidden states was automatically validated; for the single-view model (i.e., LDCRF), we varied it from 2 to 4; for multi-view models, we varied it from 4 (2 per view) to 8 (4 per view). The regularization coefficient was also automatically validated with values 0 and 10^k, k=[-4:2:4].

5.3. Results and Discussion

Table 2 shows classification accuracy on the ArmGesture dataset. For comparison, we include the results reported in [18, 20]. The results in MM-HCRF, S-KDR-SVM, LHCRF, and CHCRF are averaged values over the 5 splits. Our MV-HCRF models (LHCRF and CHCRF) outperformed all other baselines. This shows that our approach more precisely captures the hidden interaction between views, e.g.,

[4] We divided the views in this way because the two views had different dynamics (i.e., scales); hand signals contained normalized probability estimates, where as body signals contained joint velocities.

[5] We did this for the direct comparison to the state-of-the-art result on this dataset [20].

Models	Accuracy (%)
HMM [18]	84.22
CRF [18]	86.03
HCRF [18]	91.64
HCRF ($\omega = 1$) [18]	93.86
MM-HCRF	93.79
S-KDR-SVM [20]	95.30
Linked HCRF	**97.65**
Coupled HCRF	**97.24**

Table 2. **Experimental results on the ArmGesture dataset.** We include the classification accuracy reported in [18, 20]. Our multi-view HCRF models (LHCRF and CHCRF) outperformed all the baselines; to the best of our knowledge, this is the best classification accuracy reported in the literature. $\omega = 1$ means that the previous and next observations are concatenated to produce an observation.

Models	Accuracy (%)
HMM	77.67
CRF	53.30
HCRF	78.00
Linked HCRF	**87.00**
Coupled HCRF	**86.00**

Table 3. **Experimental results on the NATOPS dataset.** Our MV-HCRF models outperformed all the baselines. The only non-latent model, i.e., CRF, performed the worst, suggesting that it is crucial to learn a model with latent variables on this dataset.

Models	Accuracy (%)
CRF	90.80
LDCRF	91.02
Linked LDCRF	**92.51**
Coupled LDCRF	**92.44**

Table 4. **Experimental results on the ArmGesture-Continuous dataset.** Our linked LDCRF outperformed the two baselines.

when the left arm is lifted (or lowered), the right arm is lowered (or lifted) (see the gestures EV and SV in Figure 3). We note that S-KDR-SVM was trained on a smaller dataset size (N=10) [20]. To the best of our knowledge, our result is the best classification accuracy reported in the literature.

Table 3 shows classification accuracy on the NATOPS dataset. All of our multi-view models outperformed the baselines. The accuracy of CRF was significantly lower than other latent models. This suggests that, on this dataset, it is crucial to learn the sub-structure of the data using latent variables. Figure 5 shows an ROC plot averaged over all 6 classes, and a confusion matrix from the result of CHCRF. As expected, most labeling errors occurred within each gesture pair (i.e., #1-#2, #3-#4, and #5-#6). We can see from the ROC plot that our multi-view models reduce both false positives and false negatives.

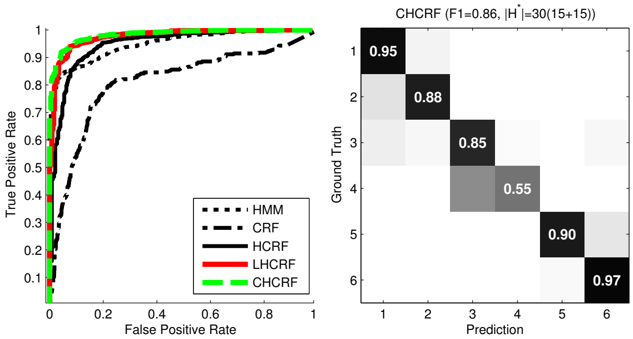


Figure 5. **ROC curve and a confusion matrix from the NATOPS experiments.** As expected, most labeling errors occurred within each gesture pair. Top of the confusion matrix shows the F1 score and the number of hidden states of CHCRF. See the text for detail.

Detailed analysis from the NATOPS experiment revealed that the multi-view models dramatically increased the per gesture pair classification accuracy; for the first and the third gesture pairs (i.e., #1-#2 and #5-#6), CHCRF achieved an accuracy of 91.5% and 93.5%, while for HCRF they were 75% and 76%. We claim that this result empirically demonstrates the benefit of our approach when each view has different dynamics (i.e., one view with velocity measures vs. another view with probability measures). The second gesture pair showed inferior performance in our multi-view models; CHCRF achieved 70%, while for HCRF it was 80%. This raises an interesting point: when one view (hand) contains almost no information, forcing the model to capture the interaction between views results in inferior performance, possibly due to the increased complexity. This observation suggests automatically learning the optimal topology of $\mathcal{E}_P$ and $\mathcal{E}_S$ could help solve this problem; we leave this as future work.

Table 4 shows per-frame classification accuracy on the ArmGesture-Continuous dataset. Consistent with our previous experimental results, the linked LDCRF outperformed the single-view LDCRF. This shows that our approach outperforms single-view discriminative models on both segmented and unsegmented action recognition tasks.

Although the linked and coupled HCRF models capture different interaction patterns, these models performed almost similarly (see Table 2 and Table 3). We believe this is due to the high sampling rate of the two datasets, making the two models almost equivalent. We plan to investigate the difference between these models using higher order Markov connections between views for the coupled models.

6. Conclusion

We introduced multi-view latent variable discriminative models that jointly learn both view-shared and view-specific sub-structures, explicitly capturing the interaction between views using disjoint sets of latent variables. We evaluated our approach using synthetic and real world data, for both segmented and unsegmented human action recognition, and demonstrated empirically that our approach successfully captures the latent interaction between views, achieving superior classification accuracy on all three human action datasets we evaluated.

Our multi-view approach has a number of clear advantages over the single-view approach. Since each view is treated independently at the input feature level, the model captures different dynamics from each view more precisely. It also explicitly models the interaction between views using disjoint sets of latent variables, thus the total number of latent variables increases only linearly in the number of views, as compared to the early fusion approach where it increases exponentially. Since the model has fewer parameters to estimate, model training requires far less training data, making our approach favorable in real world applications.

In the future, we plan to extend our models to work with data where the views are not explicitly defined. Currently we assume a priori knowledge about the data and manually define the views. However, in many real world tasks the views are not explicitly defined. In this case, we may be able to perform independent component analysis [9] or data clustering [10] to automatically learn the optimal view configuration in an unsupervised manner. We look forward to exploring this in the future.

Acknowledgments

This work was funded by the Office of Naval Research Science of Autonomy program #N000140910625, the National Science Foundation #IIS-1018055, and the U.S. Army Research, Development, and Engineering Command (RDECOM).

References

[1] P. K. Atrey, M. A. Hossain, A. El-Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Syst.*, 16(6):345–379, 2010. 1

[2] A. Blum and T. M. Mitchell. Combining labeled and unlabeled sata with co-training. In *COLT*, 1998. 2

[3] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *CVPR*, pages 994–999, 1997. 1, 2

[4] N. Chen, J. Zhu, and E. Xing. Predictive subspace learning for multi-view data: a large margin approach. In *NIPS*, pages 361–369, 2010. 2

[5] C. M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in presence of view disagreement. In *UAI*, 2008. 1

[6] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999. 4

[7] T. M. T. Do and T. Artières. Large margin training for hidden markov models with partially observed states. In *ICML*, page 34, 2009. 4

[8] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comp.*, 16(12):2639–2664, 2004. 2

[9] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley-Interscience, 2001. 8

[10] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, 1999. 8

[11] Y. Jia, M. Salzmann, and T. Darrell. Factorized latent spaces with structured sparsity. In *NIPS*, pages 982–990, 2010. 1, 2

[12] M.-A. Krogel and T. Scheffer. Multi-relational learning, text mining, and semi-supervised learning for functional genomics. *Machine Learning*, 57(1-2):61–81, 2004. 2

[13] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001. 6

[14] G. R. G. Lanckriet, N. Cristianini, P. L. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, 2004. 2

[15] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR*, 2007. 1, 2, 3, 6

[16] K. P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UCB, 2002. 1, 2, 4

[17] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI*, pages 467–475, 1999. 4

[18] A. Quattoni, S. B. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *T-PAMI*, 29(10):1848–1852, 2007. 1, 2, 3, 4, 5, 6, 7

[19] L. R. Rabiner. *A tutorial on hidden Markov models and selected applications in speech recognition*, pages 267–296. Morgan Kaufmann Publishers Inc., 1990. 4, 6

[20] A. Shyr, R. Urtasun, and M. I. Jordan. Sufficient dimension reduction for visual sequence classification. In *CVPR*, pages 3610–3617, 2010. 5, 6, 7

[21] A. F. M. Smith and G. O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Roy. Stat. Soc. Series B*, 55:2–23, 1993. 4

[22] Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *FG*, pages 500–506, 2011. 1, 5, 6

[23] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *JMLR*, 7:1531–1565, 2006. 2

[24] C. A. Sutton, K. Rohanimanesh, and A. McCallum. Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data. In *ICML*, 2004. 2

[25] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. 1

[26] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In *CVPR*, pages 872–879, 2009. 4, 6

[27] L. Wu, S. L. Oviatt, and P. R. Cohen. Multimodal integration - a statistical view. *IEEE Transactions on Multimedia*, 1(4):334–341, 1999. 1, 2