Learning Structural Models in Multiple Projection Spaces

Roman Filipovych and Eraldo Ribeiro

Computer Vision and Bio-Inspired Computing Laboratory
Department of Computer Sciences
Florida Institute of Technology
Melbourne, FL 32901, USA
{rfilipov,eribeiro}@fit.edu

Abstract. In this paper, we present an algorithm for learning structures of Bayesian models in multiple projection spaces. We assume that a visual phenomenon can be projected on a set of spaces that share a common subspace. We propose that models of individual projections can be related through probability distributions over the shared subspace. We develop a learning method that estimates simultaneously the structure and parameters of an integrated model of the target phenomena. This integrated model combines information from all individual projections. The model learning procedure is accomplished by maximizing the Bayesian Information Criterion within the setup of the Expectation-Maximization algorithm. Finally, we show how the method can be applied to the problem of learning and recognizing human motions.

Keywords: constrained Bayesian trees, substructure discovery, learning, E.M. algorithm, motion recognition.

1 Introduction

In computer vision, recognition (e.g., human-motion, objects, faces, textures) is usually accomplished based on models learned from measurements performed in projection spaces (e.g., edge-maps, 3D range data, intensity images). Unfortunately, important information about the observed phenomenon is lost during projection space creation (e.g., 3D to 2D mapping, noisy sensors, single-view camera, etc). To overcome this problem, it would be beneficial to combine models from individual projections into a single model. Neuroscientific [3, 4] evidence suggests that recognition can be more effective by combining different types of visual information. The combination of multiple visual sources can help solve problems such as image segmentation [13, 16], edge detection [11], object recognition [15, 9], and action analysis [14, 17]. However, it is not clear how data from different projection spaces can be used to build a unified structural model of a visual phenomenon.

In this paper, we propose an algorithm for learning probabilistic structural models of visual phenomena in multiple projection spaces. Our method's key assumptions are: (1) Projection spaces can be decomposed into a location subspace

and a measurement subspace, and (2) The location subspace is shared among all projections, while the measurement subspace is unique to each projection. Our main contributions are twofold. First, we show how multiple projection models can be combined into a single integrated model. Secondly, we provide an Expectation-Maximization (E.M.) algorithm for estimating both the structure and parameters of the integrated model in the presence of augmented projection spaces (i.e., location subspaces of lower dimension require augmentation for subspace sharing). Our learning approach has three main components: (1) Constrained maximization of the expectation in the E.M. algorithm; (2) Parameter initialization guided by models in non-augmented spaces; and (3) Structure selection based on the partial models' approximate contributions to the Bayesian Information Criterion (BIC). Finally, we perform a set of validating experiments on synthetic data, followed by a classification experiment on human motion data.

Our focus is on model-based recognition approaches. Model-based approaches include higher-level knowledge about the data using a previously learned model. A number of related approaches for object recognition [9] and action analysis [14, 8] use graphical models to describe both the overall structure and appearance of visual phenomena. However, model selection and the availability of prior information are issues still to be addressed.

An important but unexplored aspect of model-based approaches is the combination of different sources of information about a visual phenomenon. Multiple types of information are often integrated using combinations of classifiers [15, 11, 13]. In these approaches, a final classifier is built using a weighted combination of individual classifiers created for every information source. For example, Landy and Kojima [11] in their edge-detection approach combined different texture cues using a weighted average based on cue reliability. Nilsback and Caputo [15] propose a cue-integration framework based on a linear combination of margin-based classifiers. Leibe *et al.* [13] use top-down segmentation to integrate multiple visual cues for object detection. In contrast, Niebles and Fei-Fei [14] propose an action-recognition approach that combines multiple feature types into a constellation of bags of features. Filipovych and Ribeiro [9] propose a part-based object model that incorporates information from multiple cues for object recognition. However, their semi-supervised approach to learning object models from segmented images does not scale to general data.

As exact learning and inference are usually intractable in real scenarios, authors often resort to approximate methods. However, approximate methods, such as the E.M. algorithm, are sensitive to parameter initialization. Moreover, correct initialization is difficult in the presence of significant levels of noise in the training data. Additionally, model structure is not always known. In this case, the learning algorithm must be able to discover both the structure and parameters of the underlying model. To accomplish this, various criteria that measure "goodness" of the specific model are often used. Among commonly used model selection criteria are the Bayesian Information Criterion (BIC) [10] and the Akaike Information Criterion (AIC) [1]. Here, a search algorithm is employed to discover the model structure that receives a high score based on the selected

criterion. However, this process may become computationally intensive as the criterion needs to be evaluated for every possible model structure.

2 Unified Model of Multiple Projection Spaces

Let V represent the space of all visually perceivable phenomena. These phenomena can be any thing perceived by our vision system (e.g., objects, motions, colors, scenes, etc). Let $\Phi_i : \mathrm{V} \rightarrow \mathrm{S}_i$ be a family of general mappings or projections from V into a set of spaces S_i, with $i = 1, \ldots, n$. Accordingly, a specific target phenomenon $\mathrm{F} \in \mathrm{V}$ can be projected onto a set of spaces $\mathrm{S}_1, ..., \mathrm{S}_n$ under the corresponding mapping Φ_i. In this paper, Φ_i is a generalization of a visual cue extraction process, while S_i is the space of all visual phenomena expressed in terms of a specific cue. For example, if F represents a human face, an edge-detection method would be the cue extraction process, and the face's edge-map would be the extracted cue in the space S_i. We are particularly interested in modeling the interplay between visual cues extracted from visual phenomena. In many computer vision applications, it is common for spaces S_i to intersect along a common subspace. By definition, this intersection is itself a subspace of S_i. We denote this shared subspace by L. For example, in a multiple cue representation of an object, L can be the subspace of image pixel coordinates shared by a number of different visual cue measurements. In the case of human activities in videos, L may represent the spatio-temporal coordinates of the measurements performed in corresponding locations. Considering this space intersection, we further assume that S_i can be decomposed into two subspaces L_i and D_i such that $\mathrm{S}_i = \mathrm{L}_i \times \mathrm{D}_i$, with $\mathrm{L}_i \subset \mathrm{L}$ and $\mathrm{D}_i \cap \mathrm{L}_i = \emptyset$. Here, subspace L is shared by all S_i while subspace D_i is specific to S_i. As expected, subspaces L_i may have lower dimensionality than the common subspace L. Additionally, we assume that values from any two subspaces D_i and D_j are independent for $i \neq j$.

In computer vision, the model of a target phenomenon can be realized based on measurements performed on projected spaces. Let M_i be a model of F in space S_i where the combination of models $\mathrm{M}_1, \ldots, \mathrm{M}_n$ explain F. Here, M_i is based solely on measurements obtained from S_i, and we will refer to M_i as partial models of phenomenon F. The level of dependence between partial models can be encoded through probability functions over the common subspace L. Assuming that the probabilities of projection spaces are conditionally independent given F, the joint probability of the phenomenon and its projections can be given by:

$$p(\mathrm{F}, \mathrm{S}_1, ..., \mathrm{S}_n) = p(\mathrm{F})p(\mathrm{S}_1, ..., \mathrm{S}_n|\mathrm{F}) = p(\mathrm{F}) \prod_i p(\mathrm{S}_i|\mathrm{M}_i) \tag{1}$$

Let $\mathrm{C}_i \in \mathrm{L}$ be the origin of partial model M_i. This origin represents the partial model's abstract location in the subspace L in a similar manner as does the center of mass of a system of particles [7]. In our model, the prior distribution in (1) describes the relationships between partial models through their origins. Assuming that the probabilities of origins are conditionally independent given corresponding models (e.g., the origin of a model describing edge-map projection

is independent of a model in the space of surface normals), and by applying Bayes' rule, we can rewrite $p(\mathrm{F})$ as:

$$p(\mathrm{F}) = p(\mathrm{C}_1, ..., \mathrm{C}_n)p(\mathrm{M}_1, ..., \mathrm{M}_n|\mathrm{C}_1, ..., \mathrm{C}_n) \;\; \propto \;\; p(\mathrm{C}_1, ..., \mathrm{C}_n) \prod_i p(\mathrm{M}_i)p(\mathrm{C}_i|\mathrm{M}_i) \tag{2}$$

By plugging (2) into (1), we obtain in the following parametrized form:

$$p(\mathrm{F}, \mathrm{S}_1, ..., \mathrm{S}_n, \boldsymbol{\theta}) \propto \underbrace{p(\mathrm{C}_1, ..., \mathrm{C}_n|\boldsymbol{\xi})}_{\substack{\text{intermodel} \\ \text{prior}}} \prod_i \underbrace{p(\mathrm{S}_i|\mathrm{M}_i, \boldsymbol{\zeta}_i)}_{\substack{\text{intramodel} \\ \text{likelihood}}} \underbrace{p(\mathrm{M}_i|\boldsymbol{\zeta}_i)}_{\substack{\text{intramodel} \\ \text{prior}}} \underbrace{p(\mathrm{C}_i|\mathrm{M}_i, \boldsymbol{\zeta}_i)}_{\substack{\text{model origin} \\ \text{likelihood}}} \tag{3}$$

where $\boldsymbol{\theta}$ is the set of parameters consisting of a subset of intramodel parameters $\boldsymbol{\zeta}_i$, and a subset of parameters $\boldsymbol{\xi}$ of the intermodel prior.

As mentioned above, $\dim(\mathrm{L}_i) \leq \dim(\mathrm{L})$ for some S_i. For instance, in the case of mappings obtained from frontal human faces, the subspace of 2D image pixel locations has lower dimensionality than the subspace of the 3D range image points. Consequently, the model in (3) requires the subspace L_i to be augmented to L. This augmentation results in extending original subspace vectors by additional coordinates $\mathrm{L}_i{}^{\mathrm{aug}} = (l^{\mathrm{aug}}_{i,m+1}, \ldots, l^{\mathrm{aug}}_{i,n})$, where $\dim(\mathrm{L}_i) = m$ and $\dim(\mathrm{L}) = n$. Figure 1(a) illustrates the subspace augmentation process

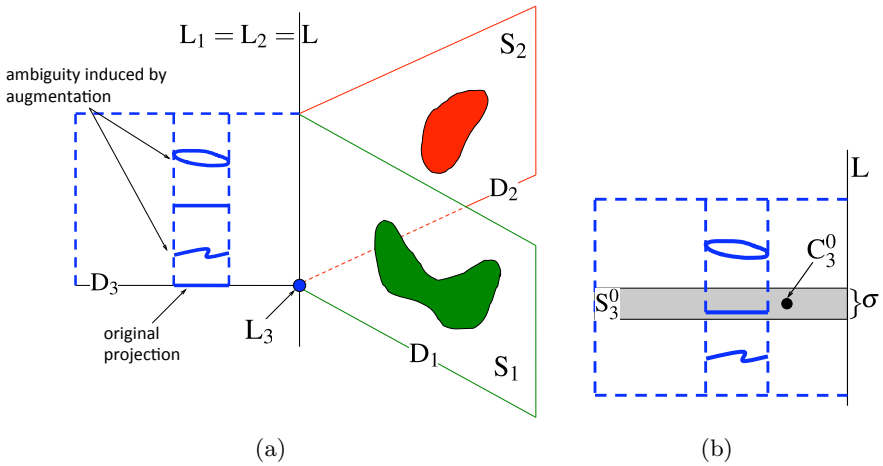

Fig. 1. (a) Subspace augmentation. (b) Reducing state space for augmented model initialization.

in which a phenomenon is projected onto three spaces S_1, S_2, and S_3, where

$\dim(\mathrm{L}_3) < \dim(\mathrm{L})$. Notice how augmentation of L_3 to L introduces ambiguity into the space $\mathrm{L} \times \mathrm{D}_3$. Space augmentation significantly increases the uncertainty of models created in these spaces. This uncertainty is reflected directly in the model learning procedure (e.g., larger probability space, increased noise levels, reduced sampling coverage, etc.). In our work, we address this issue by separately modeling augmented and non-augmented spaces. Accordingly, we denote by $\mathrm{M}_{\mathrm{U}_1}, \ldots, \mathrm{M}_{\mathrm{U}_p}$ the set of models for projection spaces that did not require augmentation, and by $\mathrm{M}_{\mathrm{K}_{p+1}}, \ldots, \mathrm{M}_{\mathrm{K}_n}$ the set of models in augmented spaces. Our learning algorithm uses models in the non-augmented spaces to initialize parameters of the augmented space models.

3 Constrained Learning in Multiple Projection Spaces

The parameters of the proposed model can be learned using the E.M. algorithm [5]. A classical E.M. procedure iterates between the E-Step:

$$p(\mathbf{Z}|\mathrm{F}, \mathrm{S}_1, ..., \mathrm{S}_n, \boldsymbol{\theta}^{old}) \tag{4}$$

and the M-Step: $\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} E_{\mathbf{Z}}\Big\{\log p(\mathbf{Z}, \mathrm{F}|\mathrm{L} \times \mathrm{D}_i, ..., \mathrm{L} \times \mathrm{D}_n, \boldsymbol{\theta}) \,|\, \mathrm{F}, \boldsymbol{\theta}^{old}\Big\}$, where original subspaces L_i have been augmented to L. Here, $\mathbf{Z}$ is a set of latent variables indicating the specific assignment of states of the model variables. However, the classical M-Step does not account for possible subspaces' augmentations. This may cause the learned model to erroneously explain data beyond the original spaces. Thus, we require that models explain data strictly within the original spaces by enforcing that $p(\mathrm{M}_i|\mathrm{L} \times \mathrm{D}_i) = p(\mathrm{M}_i|\mathrm{L}_i \times \mathrm{D}_i)$, and maximize the expectation over original subspaces:

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \left[\max_{\mathrm{L}_1^{\mathrm{aug}}, ..., \mathrm{L}_n^{\mathrm{aug}}} E_{\mathbf{Z}}\Big\{\log p(\mathbf{Z}, \mathrm{F}|\mathrm{L} \times \mathrm{D}_i, ..., \mathrm{L} \times \mathrm{D}_n, \boldsymbol{\theta})|\mathrm{F}, \boldsymbol{\theta}^{old}\Big\} \right] \tag{5}$$

Here, $\mathrm{L}_i^{\mathrm{aug}} = \emptyset$ if space S_i was not augmented.

3.1 Initialization

The above E.M. algorithm is sensitive to initialization. Better initialization can be obtained in non-augmented spaces as they usually have lower uncertainty level than the augmented ones. We propose to use non-augmented space models to initialize model parameters in augmented spaces. This motivates the following form of the intermodel prior (we assume that augmented projection models are conditionally independent given the non-augmented space models):

$$\begin{aligned} p(\mathrm{C}_1, ..., \mathrm{C}_n|\boldsymbol{\xi}) &= p(\mathrm{C}_{\mathrm{U}_1}, ..., \mathrm{C}_{\mathrm{U}_p}|\boldsymbol{\xi}_{\mathrm{U}})\, p(\mathrm{C}_{\mathrm{K}_{p+1}}, ..., \mathrm{C}_{\mathrm{K}_n}|\mathrm{C}_{\mathrm{U}_1}, ..., \mathrm{C}_{\mathrm{U}_p}, \boldsymbol{\xi}_{\mathrm{K}}) \\ &= p(\mathrm{C}_{\mathrm{U}_1}, ..., \mathrm{C}_{\mathrm{U}_p}|\boldsymbol{\xi}_{\mathrm{U}}) \prod_{i=p+1}^{n} p(\mathrm{C}_{\mathrm{K}_i}|\mathrm{C}_{\mathrm{U}_1}, ..., \mathrm{C}_{\mathrm{U}_p}, \boldsymbol{\xi}_{\mathrm{K}_i}) \end{aligned} \tag{6}$$

The initialization of model parameters can now be performed as follows:

Step 1: Initialize models in non-augmented spaces. Use the standard E.M. to initialize parameters of non-augmented space models, i.e., evaluate $p(\mathbf{Z}_{\mathrm{U}}|\mathrm{F}, \mathrm{S}_{\mathrm{U}_1}, ..., \mathrm{S}_{\mathrm{U}_p}, \boldsymbol{\theta}_{\mathrm{U}}^{old})$, and obtain:
$\boldsymbol{\theta}_{\mathrm{U}}^{new} = \arg\max_{\boldsymbol{\theta}_{\mathrm{U}}} E_{\mathbf{Z}_{\mathrm{U}}}\Big\{\log p(\mathbf{Z}_{\mathrm{U}}, \mathrm{F}|\mathrm{S}_{\mathrm{U}_1}, ..., \mathrm{S}_{\mathrm{U}_p}, \boldsymbol{\theta}_{\mathrm{U}})|\mathrm{F}, \boldsymbol{\theta}_{\mathrm{U}}^{old}\Big\}$, $\boldsymbol{\theta}_{\mathrm{U}} = \{\boldsymbol{\zeta}_{\mathrm{U}}, \boldsymbol{\xi}_{\mathrm{U}}\}$.

Step 2: Initialize global model parameters. Let $\boldsymbol{\xi}_{\mathrm{K}_i}^0$ be the initial values of the intermodel parameters of conditional probabilities in (6). Estimated model origins can be obtained by: $\mathrm{C}_{\mathrm{K}_i}^0 = \arg\max_{\mathrm{C}_{\mathrm{K}_i}} p(\mathrm{C}_{\mathrm{K}_i}|\mathrm{C}_{\mathrm{U}_1}, ..., \mathrm{C}_{\mathrm{U}_p}, \boldsymbol{\xi}_{\mathrm{K}_i}^0)$, where $\mathrm{C}_{\mathrm{K}_i}^0 = (l_{\mathrm{K}_i,1}^0, ..., l_{\mathrm{K}_i,n}^0) \in \mathrm{L}$. For every estimated origin $\mathrm{C}_{\mathrm{K}_i}^0$, we select a subspace $\mathrm{S}_{\mathrm{K}_i}^0 \subseteq \mathrm{S}_{\mathrm{K}_i}$ such that corresponding augmented coordinates are close to the augmenting coordinates $(l_{\mathrm{K}_i,m+1}^0, ..., l_{\mathrm{K}_i,n}^0)$ of origin $\mathrm{C}_{\mathrm{K}_i}^0$. This process is represented in Figure 1.(b) and can be described by:

$$\mathrm{S}_{\mathrm{K}_i}^0 = \big\{\mathbf{v} \in \mathrm{L} \times \mathrm{D}_{\mathrm{K}_i} \;\big|\; P(\mathbf{v}|l_{\mathrm{K}_i,m+1}^0, ..., l_{\mathrm{K}_i,n}^0,\ \sigma) > \mathrm{const}\big\} \tag{7}$$

Here, σ are acceptable deviation from the values $(l_{\mathrm{K}_i,m+1}^0, ..., l_{\mathrm{K}_i,n}^0)$. Initial values for non-augmented spaces' model parameters can be obtained by (E-Step):

$$p(\mathbf{Z}_{\mathrm{K}_i}|\mathrm{F}, \mathrm{S}_{\mathrm{K}_{p+1}}^0, ..., \mathrm{S}_{\mathrm{K}_n}^0, \boldsymbol{\zeta}_{\mathrm{K}_i}) \tag{8}$$

and M-Step:

$$\boldsymbol{\zeta}_{\mathrm{K}_i}^{new} = \arg\max_{\boldsymbol{\zeta}_{\mathrm{K}_i}} E_{\mathbf{Z}_{\mathrm{K}_i}}\Big\{\log p(\mathbf{Z}_{\mathrm{K}_i}, \mathrm{F}|\mathrm{S}_{\mathrm{K}_i}^0, \boldsymbol{\zeta}_{\mathrm{K}_i})\ |\mathrm{F}, \boldsymbol{\zeta}_{\mathrm{K}_i}^{old}\Big\} \tag{9}$$

Given the estimated $\boldsymbol{\zeta}_{\mathrm{K}_i}$ and subspaces $\mathrm{S}_{\mathrm{K}_i}^0$, the origins $\mathrm{C}_{\mathrm{K}_i}^0$ can now be reestimated as $\mathrm{C}_{\mathrm{K}_i}^0 = \arg\max_{\mathrm{C}_{\mathrm{K}_i}} p(\mathrm{C}_{\mathrm{K}_i}|\mathrm{M}_{\mathrm{K}_i})$. This in turn allows us to reestimate $\boldsymbol{\xi}_{\mathrm{K}_i}^0$ in (6).

Step 3: Reestimate model parameters. Given the model parameters obtained in Steps 1 and 2, we re-estimate global model parameters using (4) and (5). However, $\boldsymbol{\xi}_{\mathrm{K}_i}^0$ in Step 2 may be far from optimal. To solve this problem, we add a number of redundant augmented space models into the global model by obtaining a set of random samples of $\boldsymbol{\xi}_{\mathrm{K}_i}$. We then gradually reduce the number of augmented projection models as described next.

3.2 Model Selection

Let $\boldsymbol{\xi}_{\mathrm{K}_i,1}^0, ..., \boldsymbol{\xi}_{\mathrm{K}_i,m}^0$ be a set of m samples obtained for the intermodel parameters $\boldsymbol{\xi}_{\mathrm{K}_i}$ in (6). For every sample $\boldsymbol{\xi}_{\mathrm{K}_i,j}^0$, we obtain the reduced space given by (7) and initial parameters of the augmented space model following (8) and (9). This results in a number of redundant augmented space models that are subsequently removed by a search for the best configuration of augmented space models using a gradient ascent search over the Bayesian Information Criterion (BIC). We

Algorithm 1: Model selection algorithm

Input: Initialized global model with redundant partial models
Output: Global Model
Update model parameters following Equations 4 and 5.
Calculate the value of the $\log p(D|\mathrm{F}^h)$ for the global model.
while $\log p(D|\mathrm{F}^h) > \mathit{Threshold}$ **do**
1. Remove the partial model with the largest value of $\log p(D|\mathrm{M}_i^h)$.
2. Update model parameters following Equations 4 and 5.

end

propose an algorithm that iteratively prunes redundant augmented space models based on their approximate contribution to the overall BIC. Formally, for an alternative global model hypothesis F^h, the BIC is given by [10] as $\log p(D|\mathrm{F}^h) = ML_{\mathrm{F}^h} + \frac{d_{\mathrm{F}^h}}{2} \log N$, where ML corresponds to the maximum likelihood (ML) configuration $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, d is the number of model variables, and N is the number of data instances in D. Let $\mathbf{Z}_{\mathrm{M}_i^h}$ be the assignment of states in the overall ML configuration associated with the variables of model M_i^h. The approximate effect of including partial model M_i^h into the global model can now be estimated as:

$$\log p(D|\mathrm{M}_i^h) = ML_{\mathrm{M}_i^h} + \frac{d_{\mathrm{M}_i^h}}{2} \log N \tag{10}$$

where $ML_{\mathrm{M}_i^h} = p(\mathbf{Z}_{\mathrm{M}_i}|\mathrm{M}_i^h, \hat{\boldsymbol{\zeta}})$, and $\hat{\boldsymbol{\zeta}}$ is the subset of parameters in $\hat{\boldsymbol{\theta}}$ associated with model M_i^h. The model selection algorithm is described by Algorithm 1.

3.3 Selecting Model Origins

We now define the specific form of the model origin term in (3) by considering that each model M_i is a Bayesian network represented by a directed acyclic graph. We select one of the graph nodes as the model's landmark $\mathbf{s}_r^{(i)}$ (Figure 2(a)). M_i's origin can be expressed through the coordinates of the graph's landmark node in subspace L_i.

4 Experimental Results

4.1 Synthetic Data

We assessed our method's validity on a synthetically generated dataset consisting of two projection spaces: $\mathrm{S}_1 = \mathrm{L}_1 \times \mathrm{D}_1$, and $\mathrm{S}_2 = \mathrm{L}_2 \times \mathrm{D}_2$, where $\mathrm{L}_1 = \mathbb{R}^3$, $\mathrm{L}_2 = \mathbb{R}^2$, and $\mathrm{D}_1 = \mathrm{D}_2 = \mathbb{R}^3$. A tree-structured BN model was created in each of the spaces such that conditional distributions encode relative distances between the parent node and its children (Figure 2(b)). Following Figure 2(b), we can perform the following factorization:

$$p(\mathrm{M}) = p(a_r)p(a_1)p(a_1|a_r)p(a_2)p(a_2|a_r)\ p(b_r)p(b_r|a_r)p(b_1)p(b_1|b_r)p(b_2)p(b_2|b_r) \tag{11}$$

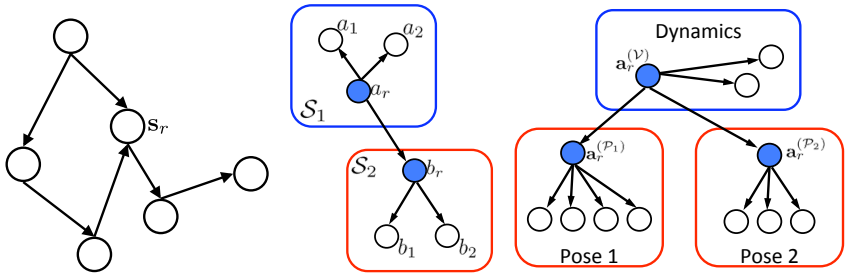

Fig. 2. (a) Bayesian network with a landmark node; (b) graph of the synthetic model; and (c) example graph of the action model.

We further simplify the model by using Gaussian densities, and assuming that priors are independent of measurements obtained in subspace L. We added noise to the sampled data points. After model learning, classification was performed with the learned model on test data containing the same amount of noise of the training data. First, we analyzed how the constrained maximization in (5) improves the performance of the algorithm. Secondly, we assessed the effect of the model selection procedure as well as performing guided initialization in our learning algorithm. Finally, we compared the performance between our learning method and the standard E.M. algorithm with randomly initialized parameters. Figure 3 summarizes our results. The plots represent evolution of the area under the ROC curve (AUC) obtained for the classification results with the amount of augmented space noise. The noise percentage in the non-augmented space is 100% for Figure 3(a), 500% for Figure 3(b), and 1500% for Figure 3(c), respectively. Results were averaged over 50 runs of the algorithm, and approximated with a least squares estimator. The results suggest that the components of our learning algorithm allow to learn a better model of the target phenomenon.

4.2 Human Motion Recognition

We now apply our learning method to the problem of learning action models from unsegmented video sequences. A human action can be projected into several projection spaces. The first projection space is represented by the video's spatio-temporal volume $\mathrm{S_V}$ and can be decomposed into two subspaces: the subspace of spatio-temporal locations $\mathrm{L_V}$, and the subspace of measurements $\mathrm{D_V}$ at corresponding spatio-temporal locations ($\mathrm{S_V} = \mathrm{L_V} \times \mathrm{D_V}$). $\mathrm{D_V}$ can be represented, for example, by spatio-temporal features [6, 12]. Alternatively, actions can also be projected on the 2D space of static pose images $\mathrm{S_P} = \mathrm{L_P} \times \mathrm{D_P}$. Here, space $\mathrm{S_P}$ represents pose information contained in a single frame, and can be decomposed into the 2D locations subspace $\mathrm{L_P}$ and a subspace of measurements at specific coordinates $\mathrm{D_P}$. In order for spaces $\mathrm{S_V}$ and $\mathrm{S_P}$ to have a common

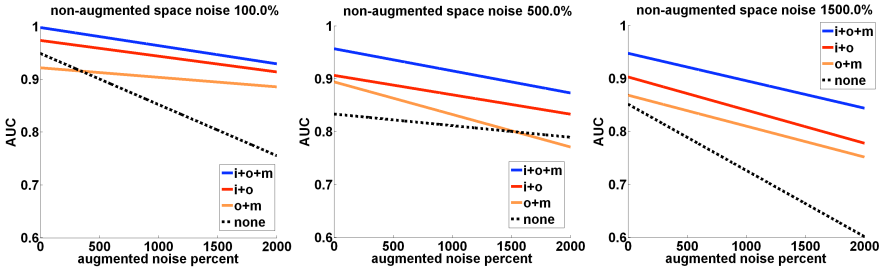

Fig. 3. AUC evolution with the the augmented space's noise amount. Plots obtained for the non-augmented space's noise: (a) 100%, (b) 500%, and (c) 1500%. Legend indicates which properties were included in the E.M. learning: (i) guided initialization; (o) model selection procedure; (m) standard E-Step of EM algorithm substituted with constrained optimization in Equation 5; and (none) for learning performed with standard EM.

subspace L, subspace $\mathrm{L_P}$ has to be augmented with an additional temporal dimension. The elements of the common subspace consequently are $(l_x, l_y, l_t) \in \mathrm{L}$, where $l_t \in \mathrm{L_P}^{aug}$. Next, we describe specific forms of models $\mathrm{M_V}$ and $\mathrm{M_P}$.

Constellation model of pose. A pose model $\mathrm{M_P}$ can be represented by a Bayesian network in the constellation framework. Consequently, pose P can be subdivided into N_P non-overlapping subregions $\mathrm{P} = \{(\mathbf{a}_1^{(\mathrm{P})}, \mathbf{d}_1^{(\mathrm{P})}), \ldots, (\mathbf{a}_{N_\mathrm{P}}^{(\mathrm{P})}, \mathbf{d}_{N_\mathrm{P}}^{(\mathrm{P})})\}$, where the components of each pair $(\mathbf{a}_j^{(i)}, \mathbf{d}_j^{(i)})$ are local appearance $\mathbf{a}$ and spatio-temporal location $\mathbf{d}$ of subregion j for the model of pose P, respectively. For simplicity, we assume that pose subregions can be arranged in a star-graph configuration in which a particular node is assigned to be a landmark node $(\mathbf{a}_r^{(\mathrm{P})}, \mathbf{d}_r^{(\mathrm{P})})$ for pose P. The distributions $p(\mathbf{d}_j^{(\mathrm{P})}|\mathbf{d}_r^{(\mathrm{P})})$ encode the relative spatial-temporal displacements of the non-landmark parts with respect to the landmark part. Additionally, if an action is projected on multiple pose spaces, a set of pose models $\{\mathrm{M_{P_1}}, ..., \mathrm{M_{P_K}}\}$ explain the set of existing pose projections.

Constellation model of motion dynamics. In a similar way, we assume a star-graph constellation model in the spatio-temporal volume. Accordingly, let $\mathrm{V} = \{(\mathbf{a}_1^{(\mathrm{V})}, \mathbf{d}_1^{(\mathrm{V})}), \ldots, (\mathbf{a}_{N_\mathrm{V}}^{(\mathrm{V})}, \mathbf{d}_{N_\mathrm{V}}^{(\mathrm{V})})\}$ be a set of N_V spatio-temporal parts in V. A node is assigned to be the dynamics model's landmark node, $(\mathbf{a}_r^{(\mathrm{V})}, \mathbf{d}_r^{(\mathrm{V})})$. Given model $\mathrm{M_V}$ in non-augmented space $\mathrm{S_V}$, and a set of models $\{\mathrm{M_{P_1}}, ..., \mathrm{M_{P_K}}\}$ in augmented spaces $\{\mathrm{S_{P_1}}, ..., \mathrm{S_{P_K}}\}$, the action model graph has the form shown in Figure 2(c). Intermodel priors can be modeled by Gaussian densities.

Discovering pose models. In the unsupervised learning of human poses from videos, the optimal number of models is unknown (i.e., the number of augmented spaces). Therefore, in the model learning algorithm (i.e., Algorithm 1), we do not enforce the requirement to have at least one model for each projection space.

Classification Results. We tested our approach on the human action dataset from [2] (Figure 4(a)). The dataset contains nine action classes performed by nine different subjects. In order to obtain the initial pose models, temporal coordinates of the origins of pose models were manually set to: $l^0_{\mathrm{P}_1,t} = -20$, $l^0_{\mathrm{P}_2,t} = -10$, $l^0_{\mathrm{P}_3,t} = 0$, $l^0_{\mathrm{P}_4,t} = 10$, $l^0_{\mathrm{P}_5,t} = 20$. The value $l^0_{\mathrm{P}_i,t}$ are the temporal displacement of pose P_i from the dynamics model's origin. Feature extraction steps were performed as in [8]. A "leave-one-out" evaluation scheme was adopted for evaluation. The confusion matrix generated by our classification results is

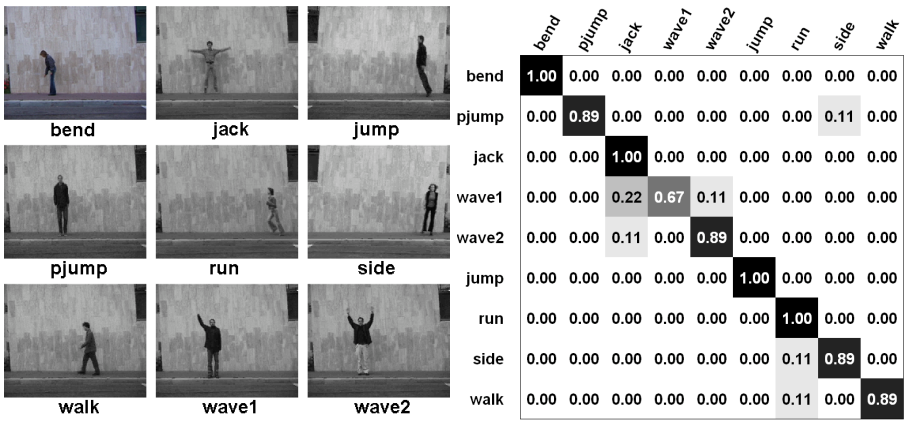

Fig. 4. Datasets in our experiments. (a) Sample frames from the human motion dataset [2]. (b) Confusion matrix for our motion classification experiment (91.34% correct classification).

.

shown in Figure 4(b), and presents a 91.34% overall recognition rate. This rate is superior to the 72.8% classification rate reported by Niebles and Fei-Fei [14].

5 Conclusions

In this paper, we presented an approach for learning both the structure and parameters of models in multiple projection spaces. Our learning algorithm incorporates the following three main components: (1) Constrained maximization of the expectation in the E.M. algorithm; (2) Model parameter initialization guided by the models in non-augmented spaces; and (3) A model structure selection procedure based on approximate contributions of partial models to the Bayesian Information Criterion. Finally, we performed a set of validating experiments, and showed that our model performs well on the human motion classification task. Future directions of investigation include studying the effect

of choosing initial intermodal parameters. Additionally, the applicability of our model to the object recognition task has to be demonstrated.

Acknowledgments This research was supported by U.S. Office of Naval Research under contract: N00014-05-1-0764.

References

1. H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
2. Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *ICPR*, pages 1395–1402, 2005.
3. Vicki Bruce, Patrick R. Green, and Mark A. Georgeson. *Visual perception : physiology, psychology, and ecology*. Hove: Psychology Press, 1990.
4. Jody C C. Culham and Kenneth F F. Valyear. Human parietal cortex in action. *Current Opinion of Neurobiology*, 2, March 2006.
5. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Stat. Soc.*, 39:1–38, 1977.
6. P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.
7. Richard Feynman, Robert Leighton, and Matthew Sands. *The Feynmen Lectures on Physics: Volume 1*, volume 1 of *The Feynman Lectures on Phsyics*. Addison-Wesley, Boston, 2nd edition edition, 1963.
8. Roman Filipovych and Eraldo Ribeiro. Combining models of pose and dynamics for human motion recognition. In *ISVC*, USA, November 2007.
9. Roman Filipovych and Eraldo Ribeiro. Probabilistic combination of visual cues for object classification. In *ISVC*, USA, November 2007.
10. David Heckerman. Bayesian networks for data mining. *Data Min. Knowl. Discov.*, 1(1):79–119, 1997.
11. Michael S. Landy and Haruyuki Kojima. Ideal cue combination for localizing texture-defined edges. *Journal of the Optical Society of America A*, 18(9):2307–2320, September 2001.
12. I. Laptev and T. Lindeberg. Space-time interest points. In *Intl. Conf. on Computer Vision*, Nice, France, October 2003.
13. B. Leibe, K. Mikolajczyk, and B. Schiele. Segmentation based multi-cue integration for object detection. In *BMVC, Edinburgh*, 2006.
14. J. C. Niebles and Li Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, Minneapolis, USA, June 2007.
15. M.E. Nilsback and B. Caputo. Cue integration through discriminative accumulation. In *CVPR*, pages II: 578–585, 2004.
16. Omer Rotem, Hayit Greenspan, and Jacob Goldberger. Combining region and edge cues for image segmentation in a probabilistic gaussian mixture framework. In *CVPR*, 2007.
17. J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A scalable approach to activity recognition based on object use. In *ICPR*, 2007.