

# Knowledge Augmented Visual Learning

Qiang Ji

Rensselaer Polytechnic Institute

[qji@ecse.rpi.edu](mailto:qji@ecse.rpi.edu)

# Motivation

- Machine learning (ML) is playing an increasingly important role in computer vision.
- As an enabler for computer vision, it allows automatically extracting pattern from the data, a significant progress over traditional hand-crafted AI-based knowledge acquisition models
- Current wisdom: powerful image features + large amount of data+ advanced learning techniques is the solution to CV ?

# Motivation (cont'd)

- Current ML methods are mostly data-driven, and they are brittle, lack of robustness, and cannot generalize well when the training data is inadequate in either quality or quantity.
- Current ML learning methods cannot lend themselves easily to exploit the readily available prior knowledge.
- Prior knowledge is essential to alleviating the problems with data and to regularize the ill-posed vision problems.

# Knowledge-Augmented Visual Learning

- Identify the related prior knowledge from different sources
- Use the Probabilistic Graphical Models (PGM) to capture and encode such knowledge systematically and automatically to produce a prior model
- Combine the prior model with image measurements (features) in a principled manner to perform visual understanding

# Sources of Knowledge

- **Permanent theoretical knowledge**
  - Various theories or principles or laws that govern the properties and behavior of the objects (e.g physics for body tracking)
  - Tend to be generic, applicable to different objects and different situations, but hard to capture
- **Subjective and experiential knowledge (expert)**
  - Knowledge gained from experience based on long time observations
  - Tend to be qualitative, inexact, and approximate
- **Circumstantial and contextual knowledge**
  - Auxiliary information or context that is available during training or testing
- **Temporary-statistical pattern-based**
  - Tend to be object, situation or database specific
  - widely used in CV.

# Methods for Knowledge Representation and Encoding

- Convert knowledge into constraints on parameters or structure of the PGM
  - Model learning can then be formulated as constrained ML/EM (either closed form or iterative )
- Numerically sample the knowledge to generate pseudo-data
  - Propose a MCMC sampling approach to efficiently explore the parameter space to acquire samples that satisfy the knowledge.
  - Encode the knowledge by the distribution of synthetic samples
  - Combine the real data with the pseudo-data to train the model

# Knowledge Representation

## MCMC Sampling

- Determine the valid range for each parameter
- Generate new sample in the valid parameter space, using the proposal distribution

$$p(\mathbf{s}^{(n)} | \mathbf{s}^{(n-1)}, \dots, \mathbf{s}^{(1)}) \propto \frac{1}{(2\pi\sigma^2)^{D/2}} - \frac{1}{n-1} \sum_{j=1}^{n-1} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left\{-\frac{\|\mathbf{s}^{(n)} - \mathbf{s}^{(j)}\|^2}{2\sigma^2}\right\}$$















- Reject samples inconsistent with the knowledge
- Repeat until enough samples are collected

The proposal distribution allows efficiently exploring the parameter space by associating high probability for unexplored regions to produce representative samples.

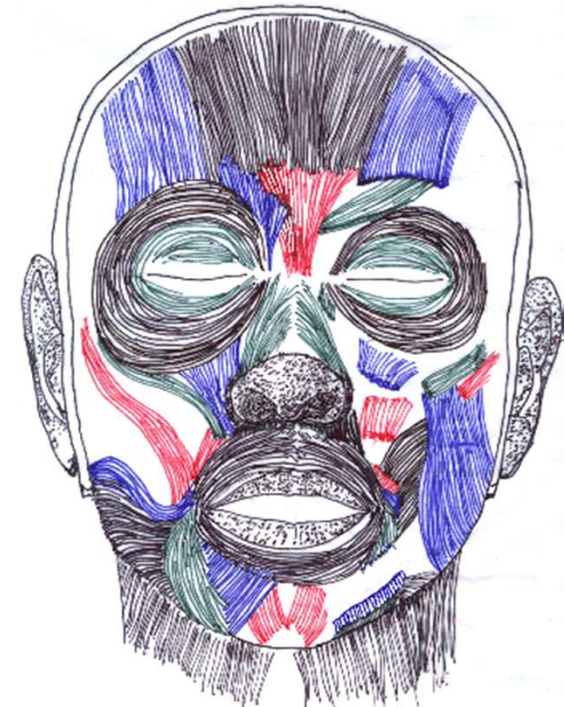
# Facial Action Recognition

(Tong and Ji, CVPR07, PAMI07, and PAMI 10)

- Facial Action Units (AUs) capture the non-rigid muscular activities that produce facial appearance changes (defined in Facial Action Coding System)
- Each AU is related to the contraction of a set of facial muscles.
- A small set of AUs can describe a large number of facial behaviors

AU1  Inner brow raiser	AU2  Outer brow raiser	AU4  Brow Lowerer	AU5  Upper lid raiser	AU6  Cheek raiser
AU7  Lid tighten	AU9  Nose wrinkle	AU12  Lip corner puller	AU15  Lip corner depressor	AU17  Chin raiser
AU23  Lip tighten	AU24  Lip presser	AU25  Lips part	AU27  Mouth stretch	

(a) A list of AUs and their interpretations



(b) Muscles underlying facial AUs

# AU Knowledge

## – Positive and negative causal influences

- Mouth stretch increases the chance of lips apart; it decreases the chance of cheek raiser and lip presser.
- Cheek raiser and lid compressor increases the chance of lip corner puller.
- Outer brow raiser increases the chance of inner brow raiser.
- Upper lid raiser increases the chance of inner brow raiser and decreases the chance of nose wrinkler.
- Lip tightener increases the chance of lip presser.
- Lip presser increases the chance of lip corner depressor and chin raiser.

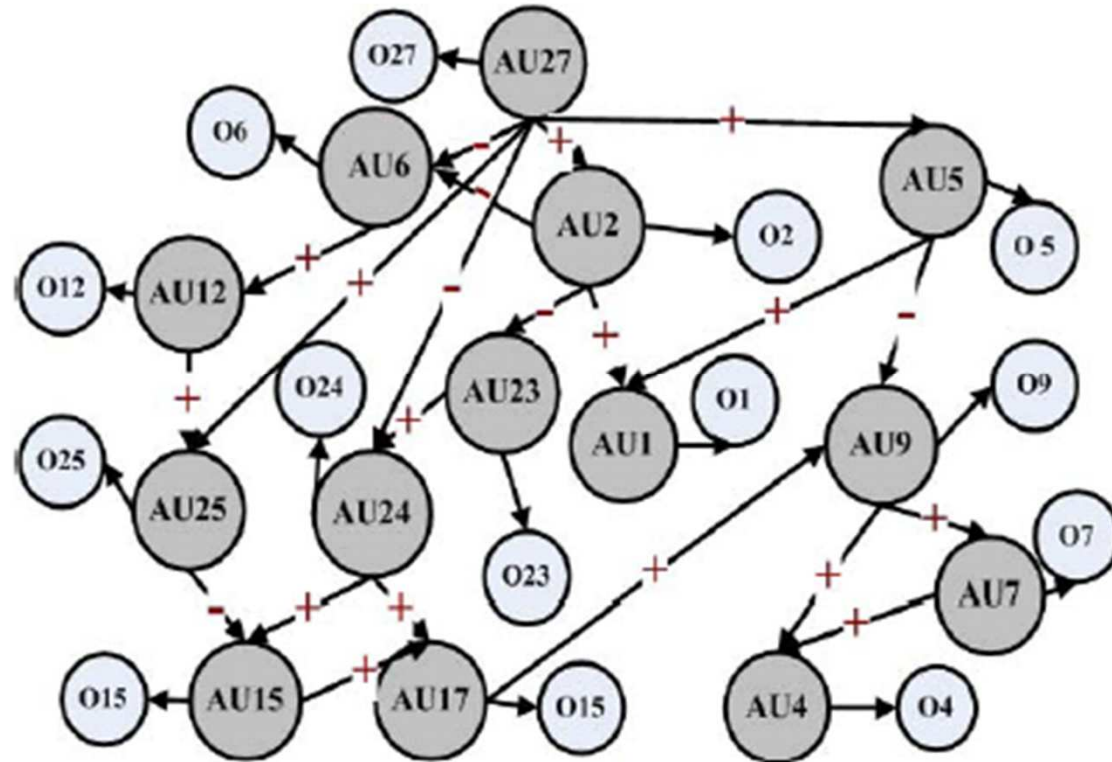
## – Group AU constraints

- Group of AUs happen together or never happen together to produce a meaningful or spontaneous expression due to underlying facial anatomy

## – Dynamic knowledge

- Each AU evolves smoothly over time
- Dynamic dependencies among AUs

# Positive and Negative Influences



For an  $AU_i$  with positive influence by its parent node  $AU_j$   $P(AU_i=1 | AU_j=1) > P(AU_i=1 | AU_j=0)$

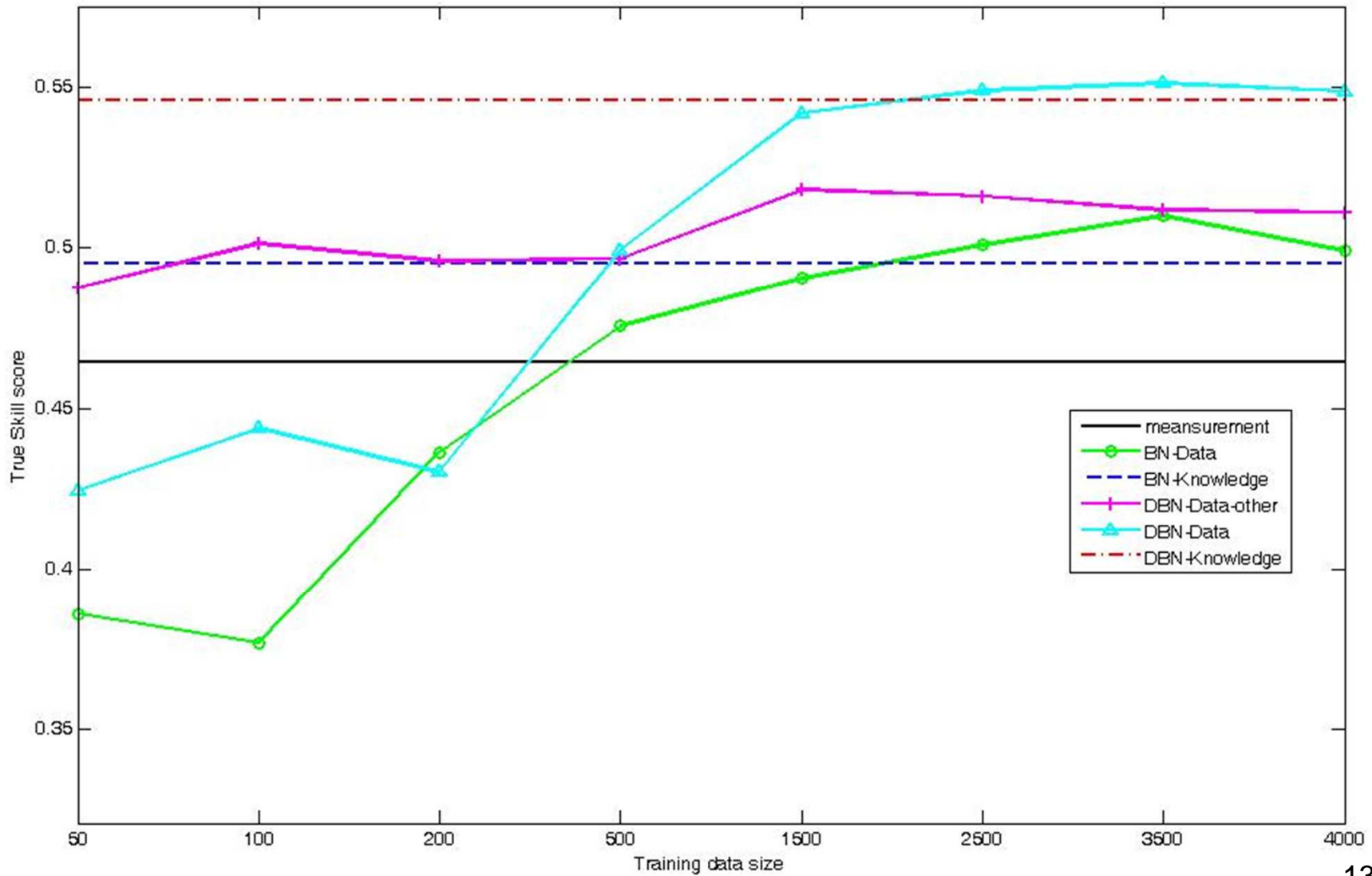
For an  $AU_i$  with negative influence by its parent node  $AU_j$   $P(AU_i=1 | AU_j=1) < P(AU_i=1 | AU_j=0)$

# AU Prior Model Learning

- Use a DBN to encode the knowledge on the relationships among AUs
- Convert the knowledge into constraints on DBN or into pseudo-data
- Learn the DBN with both pseudo and real data under constraints

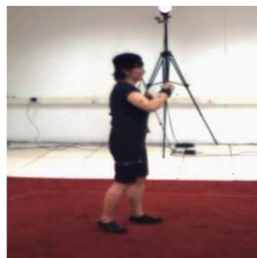


# AU Recognition Results

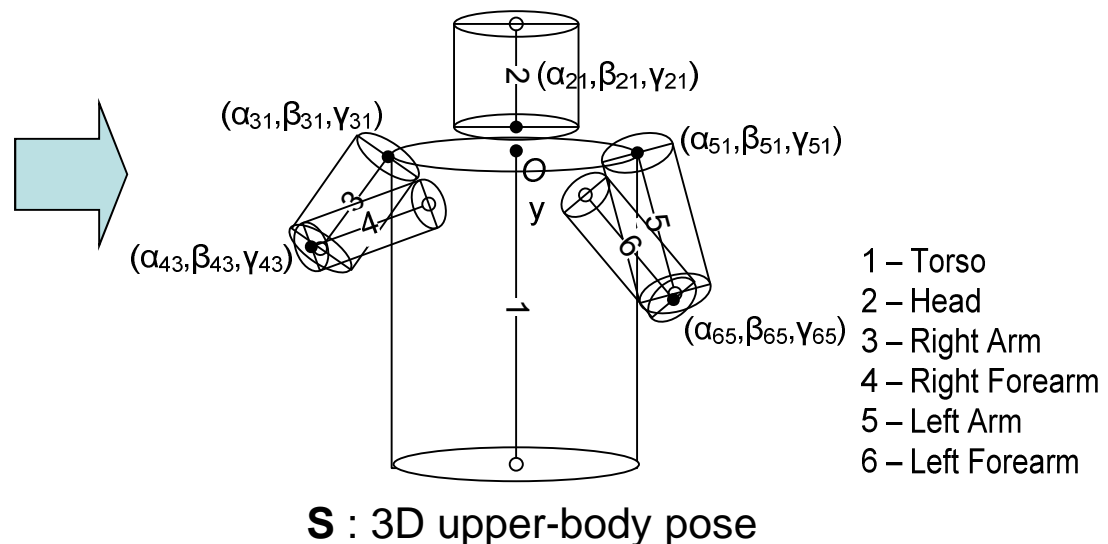


# Human Body Tracking

- **Goal:** Recover the 3D upper-body pose  $s$  given the image observation  $o$ .



$O$ : Image observation from multiple views



- The pose state is represented as the joint angles among the six rigid body parts:

$$s = (\alpha_{21}, \beta_{21}, \gamma_{21}, \alpha_{31}, \beta_{31}, \gamma_{31}, \alpha_{43}, \beta_{43}, \gamma_{43}, \alpha_{51}, \beta_{51}, \gamma_{51}, \alpha_{65}, \beta_{65}, \gamma_{65})^T$$

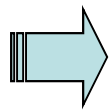
# Our Approach

- Bayesian Approach
  - Pose estimation is interpreted as the maximization of the posterior probability:  $s^* = \max_s p(s|o)$  .
  - Based on Bayes rule, the posterior can be factorized as

$$p(s|o) \propto p(o|s) \cdot p(s)$$

Image likelihood

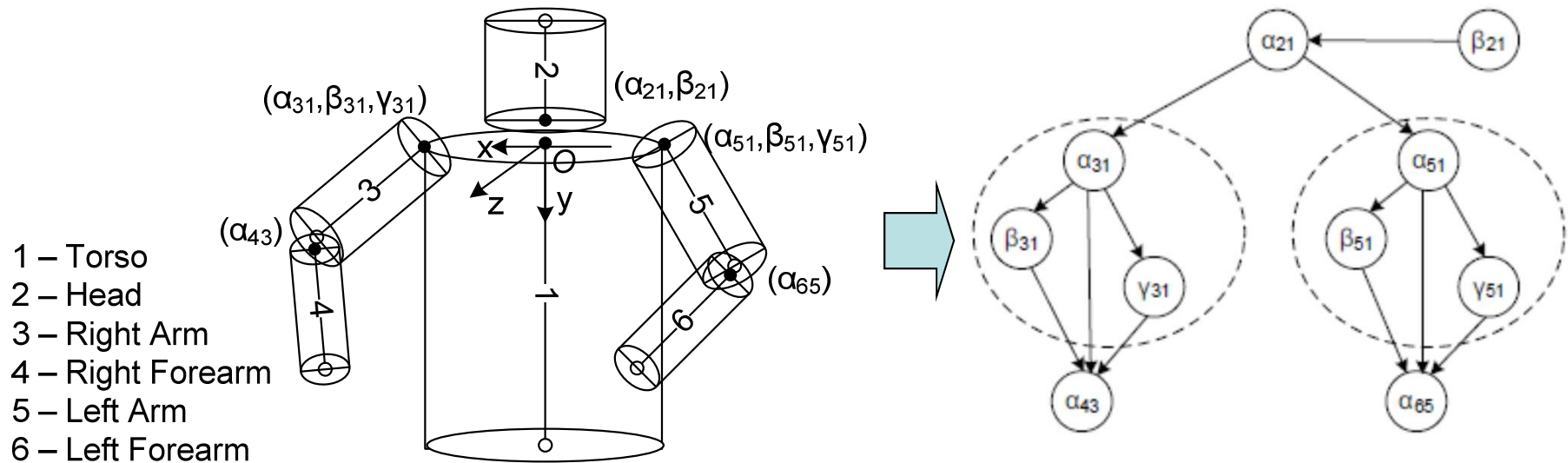
Prior model of the body pose



A good prior model can handle the uncertainty and ambiguity of the image observation

# Human Body Pose Prior Model

- We construct a Bayesian Network (BN) to model the prior probability of upper body pose.



- Node** :  $x_i$  represent the joint angle.
- Link** :  $x_j \rightarrow x_i$  represent the probabilistic relationship (mixture of Gaussians) :

$$p(x_i|x_j) = \sum_m w_m \mathcal{N}_m(\mu_m + \mathbf{W} \mathbf{x}_j, \sigma_m^2)$$

- Probability of body pose** :  $p(x_1, \dots, x_N) = \prod_i p(x_i|Pa(x_i))$

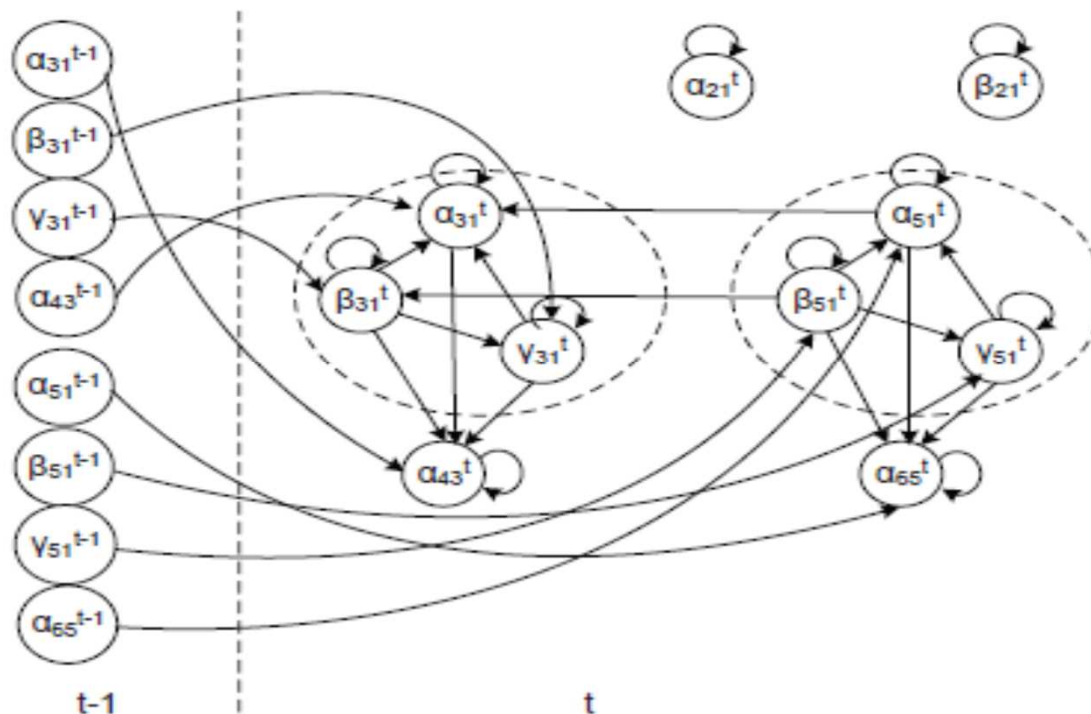
# Human Body Knowledge

- Anatomical Constraints
  - Restrict body structure based on anatomy.
    - Connectivity, kinesiology, symmetric, etc.
- Biomechanics Constraints
  - Restrict the body joint angle ranges.
- Physical Constraints
  - Exclude the physically infeasible pose
    - Non-penetrating constraint
- Dynamics Constraints
  - Restrict the body movement
    - movement speed and movement smoothness

# Knowledge-driven Model Learning

- Using the pseudo-data and constraints, learn a DBN by maximizing the score of the DBN structure (B), given pseudo data (D):

$$Score(B) = P(B) + p(D | \theta_B, B) - \frac{d}{2} \log(K)$$



# Body Tracking Experiment

- Comparison with Model from Training Data.

	Walk	Jog	Throw-Catch	Gestures	Box	Avg
Baseline	107.97	115.76	74.30	58.93	95.40	90.47

Table 1. Result of baseline system (particle filter) on 5 test sequences.

	Walk	Jog	Throw-Catch	Gestures	Box	Avg
$B_{Walk}$	60.67	168.22	51.52	95.95	121.38	99.54
$B_{Jog}$	125.50	53.48	59.24	125.99	112.03	95.24
$B_{ThrowCatch}$	109.47	167.75	42.64	57.47	114.57	98.38
$B_{Gestures}$	156.12	207.64	71.63	41.98	149.62	125.39
$B_{Box}$	143.76	173.6	51.36	63.05	68.27	100.01
$B_{HumanEva}$	46.23	86.15	45.95	43.71	81.99	60.80
$B_{CMU}$	103.26	160.45	43.56	50.63	67.11	85.00
$B_C$	67.07	70.79	45.83	48.20	74.09	61.19
$DBN_C$	67.29	75.29	44.63	44.12	66.11	59.49

Table 2. Results of different models .

**BN\_Activity** is learned from specific activity. **BN\_HumanEva** is learned from 5 activities.

**BN\_CMU** is learned from CMU database. **BN\_C** is learned from Constraints.

# Conclusions

- Knowledge is a crucial component of visual understanding, and that the long-term success of computer vision requires a union of domain knowledge and the data.
- We advocate for a hybrid approach for machine learning, whereby both knowledge and data can be integrated to result in a robust and generalizable learning.
- We propose to systemically identify related knowledge from different sources that govern the functions, properties, and behaviors of the objects being studied
- We propose to use the probabilistic graphical models to automatically and systematically capture the related knowledge and to combine with image measurements.