Action Attribute Detection from Sports Videos with Contextual Constraints

Xiaodong Yu[1], Ching Lik Teo[2], Yezhou Yang[2], Cornelia Fermüller[2], Yiannis Aloimonos[2]

[1] Comcast Corporation, Washington DC, USA, [2] University of Maryland, College Park, MD, USA

INTRODUCTION

Action Attributes:

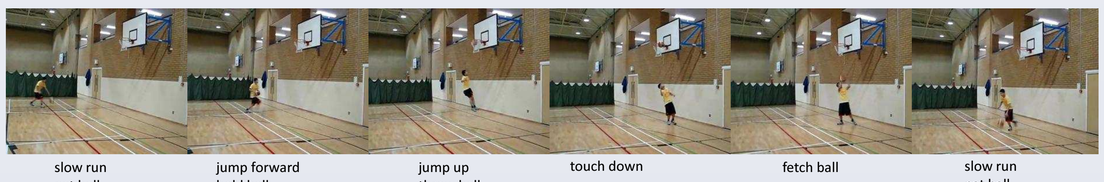
- Atomic components of action classes
  - motion patterns of human limbs and torso
- Contextual components of action classes
  - objects and scenes involved in the action
- Non-semantic attributes
  - data-driven attributes

Our Goal:

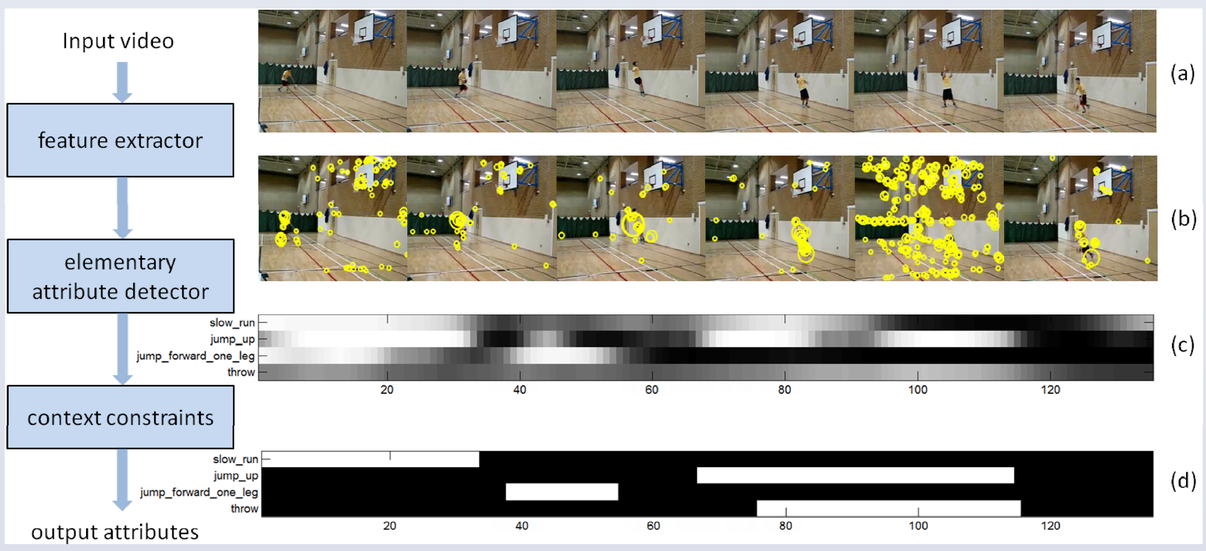
Detect action attributes and their temporal structure from sports videos

Example:

"At beginning, the athlete runs slowly while patting the ball for half a second; then he jumps forward while holding the ball for one second; then he jumps up and throws the ball (into the basket); after a touch down he fetches the ball and then runs slowly while patting the ball until the end of the video"



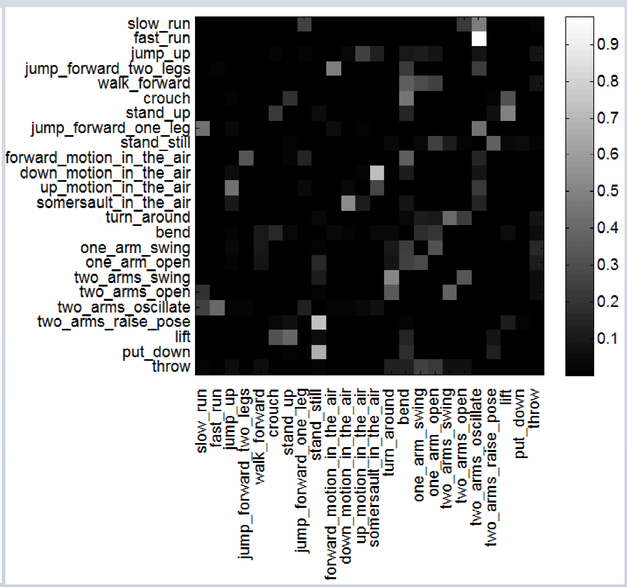
System Overview



OLYMPIC SPORTS DATASET

- 16 actions
- 20 videos for each action (15 for training, 5 for testing)
- 24 attributes
  - 9 leg motion patterns
  - 6 arm motion patters
  - 6 whole body motion patterns
  - 3 human-object interaction patterns

action-attribute matrix



attribute semantic context

ALGORITHM

Low-level Feature Extraction:

- HoG and HoF extracted at STIPs within human bounding box
- All features are quantized into one of the 400 visual words

Elementary Attribute Detectors:

- SVM with X^2 kernel

Incorporating Contextual Constraints:

- Temporal contexts
- Semantic contexts

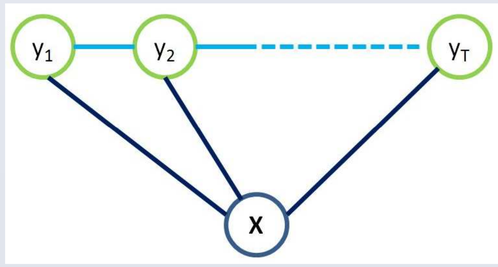
Factorial CRF Model (FCRF):

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \left(\sum_{t=1}^{T} \sum_{a=1}^{A} \Upsilon_t(y_{t,a}, \mathbf{X}) \right) \left(\sum_{t=1}^{T-1} \sum_{a=1}^{A} \Psi_t(y_{t,a}, y_{t+1,a}, \mathbf{X}) \right) \left(\sum_{t=1}^{T} \sum_{a,b \in \{1,\ldots,A\}, a \neq b} \Phi_t(y_{t,a}, y_{t,b}, \mathbf{X}) \right)$$

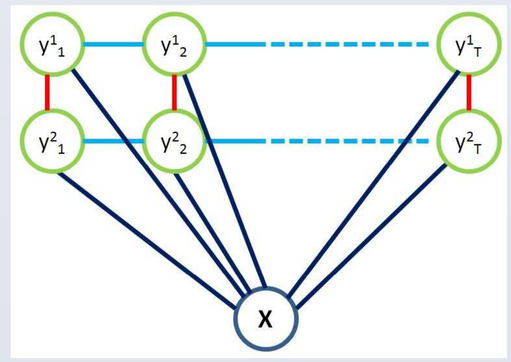
$$\exp\left(\sum_k \lambda_k f_k(y_{t,a}, \mathbf{X}) \right) \quad \text{local potential} \quad \mathbb{I}[y_{t,a} = m] \log N_a(\mathbf{x}_{t-j})$$

$$\exp\left(\sum_k \lambda_k f_k(y_{t,a}, y_{t,b}, \mathbf{X}) \right) \quad \text{temporal contexts} \quad \mathbb{I}[y_{t,a} = m \wedge y_{t+1,a} = n]$$

$$\exp\left(\sum_k \lambda_k f_k(y_{t,a}, y_{t+1,a}, \mathbf{X}) \right) \quad \text{semantic contexts} \quad \mathbb{I}[y_{t,a} = m \wedge y_{t,b} = n] \phi(a,b)$$



Linear CRF (LCRF, temporal context only)



Factorial CRF (FCRF, both temporal and contextual contexts)

Learning Model Parameters

Objective function with L2 regularization

$$L_r(\Lambda) = \sum_{n=1}^{N} \log P(\mathbf{Y}^n|\mathbf{X}^n, \Lambda) - \frac{1}{2\sigma^2} \| \Lambda \|^2$$

log likelihood of training data; L2 regularization

Stochastic gradient ascent

$$\lambda_k \leftarrow \lambda_k + \alpha \left(\sum_t f_k(\mathbf{Y}^n_{t,c}, \mathbf{X}^n) - \sum_t \sum_{\mathbf{Y}_{t,c}} p(\mathbf{Y}_{t,c}|\mathbf{X}^n) f_k(\mathbf{Y}^n_{t,c}, \mathbf{X}^n) - \frac{\lambda_k}{\sigma^2} \right)$$

gradient for training sequence X^n

RESULTS