Actionness Ranking with Lattice Conditional Ordinal Random Fields

Wei Chen, Caiming Xiong, Ran Xu and Jason J. Corso

Department of Computer Science and Engineering - SUNY at Buffalo, Buffalo, NY

Abstract: Recognizing specific actions in video clips has been the main focus of current computer vision community. We move in a new, more general direction and ask the critical fundamental question: What is action, how is action different from motion, and in a given image or video where is the action? The philosophical and visual characteristics of action lead us to define actionness: intentional bodily movement of biological agents (people, animals). In this paper, we propose the lattice conditional ordinal random field model that incorporates local evidence as well as neighboring order agreement to solve the general problem.

Action Recognition Status

Action Recognition Representation

The CV community has developed rich representations for action in video
- local action features with bags-of-words framework
--- spatio-temporal interest points,
--- trajectory-based representation,
--- motion interchange patterns,
- semantic action representation
--- action bank, ...

Example of HMDB51

Example of UCF Sports

In all these methods, we can find two points:

- What is an action?

The very notation of action has not been carefully defined or explicitly studied. Instead, action is defined implicitly by examples in a dataset.

- Action = Motion?

Motion feature is the dominant part of action representation. What is the difference between action and motion?

Based on the discussion, we propose the notation of actionness to answer these two questions.

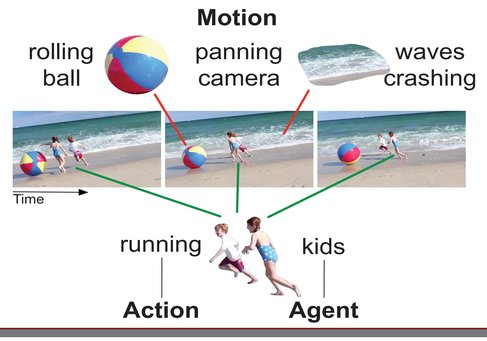
Actionness : What is an Action?

Action from the viewpoint of the philosophy

There are four aspects to define action:
- action is what an **agent** can do;
- action requires an **intention**;
- action requires a **bodily movement** guided by an agent or agents;
- action leads to **sides-effects**.

Actionness

We define actionness as intentional bodily movement of biological agents.
It is a subclass of general motion and a direct presentation of action.
It provides a non-specific definition for action.
Here, we formulate the goal of ranking image/ video regions according to their actionness, or the degree to which an agent is doing intentional bodily movement within them.



Lattice Conditional Ordinal Random Field

Problem Statement

Given an image/video data *V*, let *R* be a partitioning of *V* with *n* regions.
The partitioning can easily be computed by rectilinear patches or cubes.
Define a predicate function λ_{ij} indicating the local actionness ordering of any two regions. Our problem is to seek a valid ordering of regions, given the image/video *V* and its partitioning *R* under a local ordinal model. However, this problem is an instance of the linear ordering problem, which is NP hard.

$$\lambda_{ij} = \begin{cases} 1 & A(r_i) > A(r_j) \\ 0 & \text{otherwise} \end{cases} \qquad \Lambda^* = \underset{\Lambda,P}{\operatorname{argmax}} \sum \phi(r_i, r_j, \lambda_{ij}) \quad \text{s.t.} \quad \lambda_{ij} \in \{0,1\}, P \cdot \Lambda = U$$

LCORF Model

In order to make the problem tractable, we relax the model to be a continuous CRF model and assign a real-valued variable for each region. So the strict ordering is a partial ordering such that $a_1 \geq a_2 \geq \cdots \geq a_n$. The relaxed model is written as

$$M(\{a_i\}_{i=1}^n | \mathcal{V}, \mathcal{R}, \alpha, \beta) = \frac{1}{Z[\mathcal{R}]} \exp\left[\sum_i \alpha f(a_i, r_i) + \sum_{i,j} \beta g(a_i, a_j, r_i, r_j)\right]$$

Partitioning and Annotation

To partition each sample and compute the lattice, we simply divide the image/ video into a rectilinear set of patches (cuboids). We developed an automatic scheme that requires one or more bounding boxes around the action region. Denote the set of bounding boxes as $\{B_j\}_{j=1}^b$. *pos*() indicates the centroid of the region or the bounding box, and *D*() is the normalized Euclidead distance. The actionness score for region is as

$$a_i = 1 - \min(D(\text{pos}[r_i], \text{pos}[B_j]))$$

Unary Term

The unary term scores the actionness for each region based on its evidence.
- A trained AdaBoost classifier
--- Appearance information;
--- Spatial/ spatio-temporal information.
- The non-parameteric generalized Hough voting.
--- deal with underlying varied appearance of actionness;
--- A codebook is learnt via k-means clustering and local appearance information.
--- The score map is computed as the mean over region hough scores
The product of these two factors contributes to the measurment of evidence. The unary function is as follows

$$f(a_i, r_i^{(q)}) = -(a_i - \hat{a}_i^{(q)})^2$$

Pairwise Term

- This term enforces an ordering locally betwen two regions based on the appeance information.
- The local order preference is then computed by a trained AdaBoost classifier on the possible neighboring relations on the lattice.

$$g(a_i, a_j, r_i, r_j) = R_{ij}(a_i - a_j) = \delta_{ij} w(v_i, v_j)(a_i - a_j)$$

The pairwise term penalizes the current actionness scores of two regions when they disagree with the predicated relationship from the classifier.

Learning and Inference

Given the training data, we estimate the parameters $\theta = (\alpha, \beta)$ by MLE.
Inference on our lattice conditional ordinal random field is straightforward.

Learning

1: Input: training data T_r, and its associated Actionness score $A = \{A_s\}_{s=1}^t$, maximal iteration $Iter$ and learning rate η
2: Output: $\log \alpha$ and β
3: **for** $i = 1$ to $Iter$ **do**
4: **for** $k = 1$ to t **do**
5: Compute $\frac{\partial L(\theta|T_r)}{\partial \log \alpha}$ and $\frac{\partial L(\theta|T_r)}{\partial \beta}$ by Eq 13
6: Update $\log \alpha = \log \alpha + \eta \frac{\partial L(\theta|T_r)}{\partial \log \alpha}$
7: Update $\beta = \beta + \eta \frac{\partial L(\theta|T_r)}{\partial \beta}$
8: **end for**
9: **end for**

$$\frac{\partial L(\theta)}{\partial \log \alpha} = \alpha \sum_s \left[\sum_i -(a_i^{(s)} - \hat{a}_i^{(s)})^2 - \frac{\partial \log Z[\mathcal{R}_s]}{\partial \alpha}\right]$$

$$\frac{\partial L(\theta)}{\partial \beta} = \sum_s^{n_t} \left[\sum_{ij} R_{i,j}^{(s)}(a_i^{(s)} - a_j^{(s)}) - \frac{\partial \log Z[\mathcal{R}_s]}{\partial \beta}\right)$$

Inference

$$\hat{A}_i^{(e)} = \frac{2\hat{a}(r_i^{(e)})\alpha + \beta\left(\sum_j R_{ij}^{(e)} - \sum_i R_{ij}^{(e)}\right)}{\alpha}$$

Experiments

Dataset: We apply the method on Stanford40, UCFSports and Hollywood1 (action happend clippers) action datasets, which includes the action bounding boxes. We split the data into training and testing data following the previous work.

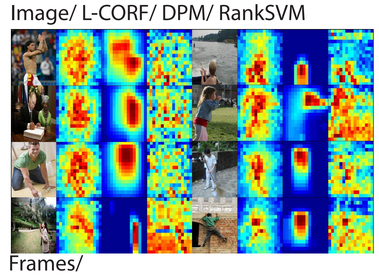
Evaluation Protocol: mean average precision (mAP) is used to judge how well the actionness score agrees with the annotation. First, we score each patch / cuboid according to the intersection over union w.r.t. groundtruth (>= 0.5, positive). Then, PR curves are generated. Each test sample will generate an AP score, which is the area under the PR curve. mAP is the average of all the test samples. In all the experiments, we divide the image and video to 16×16 grids in space. For video data, the cuboid lasts 4 frames.

Quantitative comparisons against baselines

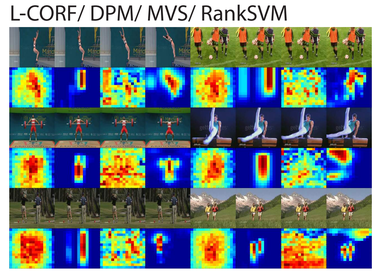
	Stanford 40	UCF Sports	HOHA
L-CORF	72.5	**60.8**	**68.5**
DPM [9]	**85.6**	54.9	60.8
RankSVM [14]	55.8	21.9	26.8
MBS [32]	-	22.8	57.4

Quantitative | Vis-Standford40 | Vis-Ucf Sports | Vis-HOHA1

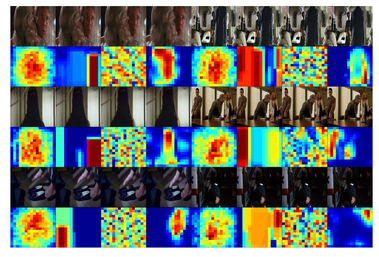
Image/ L-CORF/ DPM/ RankSVM

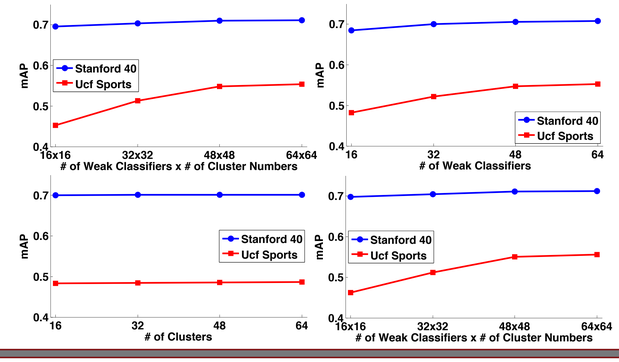


Frames/ L-CORF/ DPM/ MVS/ RankSVM



Frames/ L-CORF/ DPM/ MVS/ RankSVM





Unary | Classifier (U)
HV (U) | L-CORF

This is the first work on class-independent action analysis and video parsing. There is still a long way to go.

Acknowledgements. This work was in part supported by NSF CAREER IIS-0845282, DARPA/ARL W911NF-10-2-0062, and ARO W911NF-11-1-0090.