Actionness Ranking with Lattice Conditional Ordinal Random Fields

Wei Chen, Caiming Xiong, Ran Xu and Jason J. Corso
Computer Science and Engineering, SUNY at Buffalo
{wchen23, cxiong, rxu2, jcorso}@buffalo.edu

Abstract

Action analysis in image and video has been attracting more and more attention in computer vision. Recognizing specific actions in video clips has been the main focus. We move in a new, more general direction in this paper 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? We study the philosophical and visual characteristics of action, which lead us to define actionness: intentional bodily movement of biological agents (people, animals). To solve the general problem, we propose the lattice conditional ordinal random field model that incorporates local evidence as well as neighboring order agreement. We implement the new model in the continuous domain and apply it to scoring actionness in both image and video datasets. Our experiments demonstrate not only that our new model can outperform the popular ranking SVM but also that indeed action is distinct from motion.

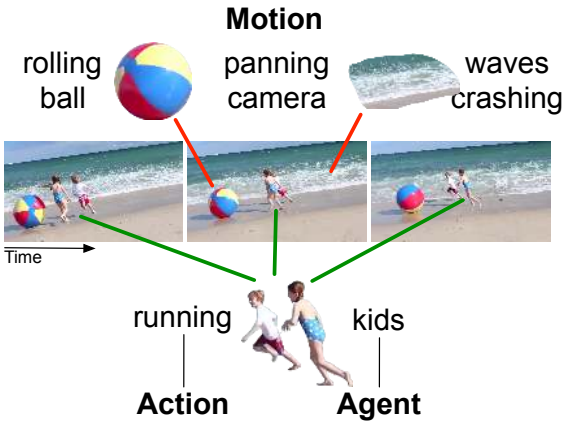


Figure 1: The key idea in our paper is to distinguish intentional action of an unknown agent (the kids in this example) from various other motions, such as the rolling ball, the crashing waves and the background motion from the panning camera. Our paper proposes a self-ordering CRF model that ranks regions of the image/video according to its agent and category independent "actionness."

1. Introduction

Human and other biological motion, such as a cat climbing a tree, present an intricate visual pattern that is of far higher complexity than most non-biological motion, such as a rolling ball or car, or simple bar and dot stimuli used in many psychophysical studies [15]. Indeed these intricate visual patterns are complex (and apparently important) enough that we humans have highly specialized parts of our brain dedicated specifically to biological motion perception (the superior temporal sulcus) [25].

Likewise, the computer vision community has achieved marked success in automatic action recognition from video. Notable examples include the introduction of local action features with bags-of-words framework [35], such as spatio-temporal interest points [21], trajectory-based representations [23, 34], and motion interchange patterns [17] and the more holistic action bank representation which embeds a video into an action space by responses of individual action detectors [31]. These methods are enabling futuristic vision applications like automatic video-to-text [5, 18] and smart classrooms [28].

However, in all of this so-called *action recognition* work in our field, the very notion of action has not been carefully defined or explicitly studied, although a hierarchy of actions and activities has been discussed [24]. Instead, action is defined implicitly by examples in a dataset. UCF Sports [29], for example, emphasizes olympic sports as action whereas HMDB51 [19] focuses more on everyday human actions such as brushing hair and hugging.

There are more explicit general notions of action [6]; most commonly, an action involves intentional biological motion. In other words, action is a specific subclass of general motion requiring an *agent* who has a particular goal or intention and is moving to achieve the goal. See Fig. 1, for example, where the two kids are chasing a beach ball near the crashing waves. There are four distinct motions in the scene but only one action: the kids running. The crashing waves, the panning camera and the rolling ball are all various motions.

Furthermore, it may be beneficial simply to detect action in a way agnostic to the specific agent carrying out the action as well as the type of action itself. For example, it has been empirically demonstrated that action classifiers perform better when they use features from foreground moving regions rather than the full video [17].

To these ends, our paper seeks to extract a rank ordering of video regions according to the degree to which they contain an action. We call this notion *actionness*. We target a rank ordering of actionness by regions rather than a direct classification of whether or not a region contains an action for two primary reasons. First, the foundational notion of action as an agent's intentional motion immediately presents a difficulty: agent (e.g, person, animal) detection remains a challenging and open problem [8]. There exist comparatively strong methods like deformable parts models [9], but the average precision remains too low for robust use (e.g., about 49.5 for person is the state of the art [11]). Ranking makes it plausible to forego agent detection and segmentation prior to actionness classification; rather, directly ranking various regions of the image/video is more robust. Second, in any given image or video there may be more than one agent performing an action. Ranking which is more likely an action is hence more informative than simple classification.

To accomplish the actionness ranking, we first propose an explicit definition of action that is based on the philosophy of action [6]. Then, we propose and implement a novel self-ordering conditional random field model that can extract the actionness ranking. Our model, called the lattice conditional ordinal random field (L-CORF), solves the linear ordering problem approximately using local features to score a given region by a generalized Hough voting framework and an unary classifier as well as pairwise relationships between neighboring regions. The pairwise term uses an AdaBoost classifier to predict the local ranking preference of two regions and penalizes the current ranking when it violates the classifier prediction. To provide an effective situation for learning and inference, we relax the discrete ordering problem in the random field to a continuous one and derive exact solutions for inference and a gradient descent method for learning.

We implement and test our model on both images and videos. In video, the agent's intentional bodily movement can be directly observed; in images, we need to rely instead on the appearance of the agent's body (i.e., the pose [39]) to infer actionness because static images have no observable motion information. In summary, action understanding benefits from motion information. However, not all motion information contributes to action understanding. Distinguishing meaningful and meaningless motion is important and will lead to better video understanding methods [12, 17].

Although we were inspired by the recent work in ranking category independent objects proposals [1, 7], our paper is the first to work on this important problem of agent and category independent actionness. Furthermore, our lattice conditional ordinal random field is an innovation on top of the conditional ordinal random field [16] that takes into account the spatial relations of regions in the lattice. Our results on both image and video actionness demonstrate the benefit of this spatial information in actionness against all baselines.

2. Actionness: What is an Action?

In this paper, we propose the notion of *actionness*, which seeks to distinguish different motions (intentional motion from general motion). Before concretely formulating the problem, we first make a definition of action suitable for computer vision, which cares more about what visual patterns an action may present than the philosophy of action.

There are four aspects to define *action* in the philosophy of action [6]: first, action is what an **agent** can do; second, action requires an **intention**; third, action requires a **bodily movement** guided by an agent or agents; and fourth, action leads to **side-effects**. For example, playing with a ball is an instance of action. A person is able to play with a ball. Doing this action needs the movement of the human body; the person moves the ball by moving his or her hands and/or feet. When a person plays with a ball, a ball movement from left to right and up to down is just a side-effect since the ball has no intention. Its movement is barely the result of the action (playing) of the person.

Above, we highlighted the key words for the four aspects of action: agent, intention, bodily movement, side-effects. Two of these are directly observable in video: agent and bodily movement (in an image, one can only observe agent pose but not the bodily movement). Intention is not directly observable but not irrelevant from a computer vision point of view: a non-biological agent, such as a bicycle can not have intention, and hence the agents we care about are people and animals. We note the discussion made in the introduction regarding the current reliability of person detectors in images. Finally, side-effects may be directly observed in images, but these would involve a complex inference even farther beyond the reliable capability of our field than person and animal detection. Therefore, we define **actionness** as intentional bodily movement of biological agents. Actionness is a subclass of general motion and a direct presentation of action.

Actionness provides a non-specific definition for action that does not rely on an absolute scale for action nor a certain type of action, which is well beyond the scope of this paper. Here, we formulate the useful goal of ranking image/video regions according to their actionness, or the degree to which an agent is *doing* intentional bodily movement within them. In the next section, we make this prob-

lem statement more concrete and then further develop a new self-ordering CRF model to perform the ranking task.

3. Lattice Conditional Ordinal Random Field

Problem Statement Given an image or video $\mathcal{V}$, let $\mathcal{R} \doteq \{r_i\}_{i=1}^n$ be a partitioning of $\mathcal{V}$ with n regions in the partitioning. Strictly, $\mathcal{R}$ is a partitioning of the pixel/voxel lattice underlying $\mathcal{V}$. The partitioning can easily be computed by rectilinear patches or cubes, which we do in this work, or by common superpixel [10] or supervoxel [37] methods, which is not the main emphasis of our work.

Given any two regions r_i and r_j, we seek an ordering of them according to their relative actionness. Although we do not seek the absolute actionness score of a region, let $A(r_i)$ denote the actionness of region r_i. Define a predicate function λ_{ij} to represent the local actionness ordering of regions r_i and r_j:

$$\lambda_{ij} = \begin{cases} 1 & A(r_i) > A(r_j) \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

And define the ordering predicate matrix Λ as the dense ordering matrix for all pairs of regions. There are 2^{n^2} possible Λ matrices but only a (still very large) fraction of these $(2^{(n^2-n)/2})$ are valid orderings of the regions. A valid ordering is captured by two properties on Λ. First, Λ is antisymmetric, i.e., $\lambda_{ij} + \lambda_{ji} = 1$ for all regions $i \neq j$. Second, there exists a permutation matrix P that will reorder the rows, λ_i, such that Λ is a strictly upper triangular binary matrix U: $P \cdot \Lambda = U$.

Our problem is thus to seek a valid Λ given the video $\mathcal{V}$ and its partitioning $\mathcal{R}$ under a local ordinal model ϕ:

$$\begin{aligned} \Lambda^* &= \arg\max_{\Lambda, P} \sum_{i,j} \phi(r_i, r_j, \lambda_{ij}) \\ \text{s.t.} \quad & \lambda_{ij} \in \{0, 1\}, P \cdot \Lambda = U \ , \end{aligned} \tag{2}$$

where the ordinal model ϕ captures the local ordering compatibility of λ_{ij} with the evidence in r_i and r_j. This local ordinal model ϕ pushes the ordering to obey the data $\mathcal{R}$ and the two constraints maintain a valid ordering.

However, the program in Eq. 2 is an instance of the linear ordering problem, which is known to be NP hard [13]. To understand this point, consider the impact of changing only one off-diagonal entry of Λ. If the entry in question relates, say, the first and second ordered regions then this is an easy and local swap with no global impact. On the other hand, if the entry relates the first and the last ordered regions then this has maximal impact (although it remains a valid ordering, every other region is now is potentially in conflict with the ordering). In general, the longer the distance in the ordering between the entry in Λ that would be swapped, the more global the impact on the ordering Λ.

3.1. The Model

Consider again the relationship upper triangular constraint: i.e., there exists a strictly upper triangular binary matrix U such that $P \cdot \Lambda = U$. This constraint implies two facts that will lead us to making an approximation to the NP hard problem. First, clearly there is a one-to-one relationship between Λ and P. Second, the sum of any row λ_i is its ordinal index $o_i = \sum_j \lambda_{ij}$. Equation 1 specifies that if $\lambda_{ij} = 1$ then $A(r_i) > A(r_j)$. So, when $A(r_i)$ is the highest actionness, then each element of row λ_i will be one (except for λ_{ii}, which is always 0). In this case, $o_i = n - 1$. A similar exercise can be conducted to demonstrate this for other ordinal indices: i.e., if region r_j is the kth index, then row λ_j will have $k - 1$ entries that are 1.

Therefore, we can reformulate our original objective to directly seek the ordinal index in a manner more readily soluble. Inspired by the recent conditional ordinal random fields [16, 30], which are defined on one-dimensional streams, we present a new model called the lattice conditional ordinal random field (L-CORF).

We propose a conditional random field model M that captures the ordering as its random variables:

$$\begin{aligned} &M_d(\{o_i\}_{i=1}^n | \mathcal{V}, \mathcal{R}, \theta) = \\ &\frac{1}{Z[\mathcal{R}]} \exp\left[\sum_i \alpha f_d(o_i, r_i) + \sum_{i,j} \beta g_d(o_i, o_j, r_i, r_j)\right], \end{aligned} \tag{3}$$

where $Z[\mathcal{R}]$ is the normalization function and $\theta = (\alpha, \beta)$ are model parameters. Recall, o_i is the ordinal index of region r_i; these indices take values from $\{1, 2, \ldots, n\}$ and satisfy a strict ordering $o_1 > o_2 > \cdots > o_n$. Functions f_d and g_d capture the unary ordinal preference and pairwise ordinal agreement, which we will make explicit below.

Satisfying the strict ordering constraint on $\{o_i\}_{i=1}^n$ and the discrete nature of this ordering make learning and inference intractable. So, we relax our model to be a continuous CRF and replace o_i with a real-valued variable a_i for each region r_i. Furthermore, we relax the strict ordering to be a partial ordering such that $a_1 \geq a_2 \geq \cdots \geq a_n$. The relaxed model is written

$$\begin{aligned} &M(\{a_i\}_{i=1}^n | \mathcal{V}, \mathcal{R}, \theta) = \\ &\frac{1}{Z[\mathcal{R}]} \exp\left[\sum_i \alpha f(a_i, r_i) + \sum_{i,j} \beta g(a_i, a_j, r_i, r_j)\right] \ . \end{aligned} \tag{4}$$

Functions f_d and g_d are continuous version of f_d and g_d.

In the following subsections, we formulate the terms of the model and derive a maximum likelihood learning method for the L-CORF.

3.2. Partitioning and Annotating

To partition each sample $\mathcal{V}$ and compute the lattice, we simply divide the image (video) into a rectilinear set of

patches (cuboids). We also need to associate an actionness evidence score with each region. Arbitrarily gathering actionness ranks / scores from humans would be prone to noise, so we instead developed an automatic scheme that requires one or more bounding boxes (or cubes in video) around the action region. Denote the set of bounding boxes as $\{B_j\}_{j=1}^{b}$. We then define the annotated actionness score a_i for region $r_i \in \mathcal{R}$ as

$$a_i = 1 - \min_j \big(D\left(\text{pos}[r_i], \text{pos}[B_j]\right)\big) \ , \tag{5}$$

where $\text{pos}[\cdot]$ indicates the centroid of the region or the bounding box and $D(\cdot)$ is the Euclidean distance. Since the size of the images/videos can differ, we normalize the distance of any two bounding boxes between 0 and 1. And this distance contributes to the actionness score computation.

3.3. Unary Term

The unary term scores the actionness for each region based on its evidence. A trained AdaBoost classifier [33] is used to measure the degree that the region includes actionness information with local appearance and spatial information. Since the underlying appearance of actionness will greatly vary, we also incorporate the non-parametric generalized Hough transform [2, 36]. Assume we have training data $T_r = \{(\mathcal{V}_s, \mathcal{R}_s, A_s)\}_{s=1}^{t}\}$ with t samples, where each A_s is the annotated actionness map (i.e., a known actionness a_i at each region in $\mathcal{R}_s$) from Sec. 3.2. Let $(\mathcal{V}_q, \mathcal{R}_q)$ denote a test image/video and its partitioning. For the generalized Hough transform, we define a scoring function $h(r_i^{(q)})$ based on the regions and their relative positions that votes on a full actionness map A_q for the test data $\mathcal{V}_q$ given a single region $r_i^{(q)}$ (the superscript (q) denotes which image/video the region is from).

To compute the voting model, we learn a codebook for each position based on appearance information. Each codebook entry c_j comprises a feature descriptor v_{c_j}; the codebook $\mathcal{C}$ is learned via k-means method. With the learned codebook $\mathcal{C}$, we define the Hough scoring function as

$$h\left(r_i^{(q)}\right) \propto \sum_j m_{c_j} p(c_j|v_{r_i}) \exp\left[-\frac{1}{\sigma} D^2(v_{r_i}, v_{c_j})\right] \ , \tag{6}$$

where m_{c_j} corresponds to an actionness map of c_j. Finally, we compute the Hough scored actionness map for the test data $\mathcal{V}_q$ as the mean over region hough scores in Eq. 6:

$$\hat{A}_q = \frac{1}{|\mathcal{R}_q|} \sum_{r_i^{(q)}} h\left(r_i^{(q)}\right) \tag{7}$$

We define the unary function as

$$f(a_i, r_i^{(q)}) = -(a_i - \hat{a}_i^{(q)})^2. \tag{8}$$

where $\hat{a}_i^{(q)}$ is the product of the Hough voting actionness score for region r_i in map $\hat{A}_q$ computed by Eq. 7 and the normalized AdaBoost classifier response.

3.4. Pairwise Term

The pairwise term enforces a certain ordering locally between two region r_i and r_j based on the features at those regions v_i and v_j. The local order preference is then computed by a trained AdaBoost classifier on the possible neighboring relations on the lattice (horizontal and vertical directions). For each neighboring relation, the classifier takes the relative actionness score for the neighboring training regions as the label (i.e., 1 if r_i has higher actionness than r_j and 0 otherwise, similar to λ_{ij} from Eq. 1). It then trains a classifier based on the features of the regions, $w(v_i, v_j)$, to predict the preferred ordered.

The pairwise term penalizes the current actionness scores of the two regions when they disagree with the predicted relationship from the AdaBoost classifier $w(v_i, v_j)$:

$$\begin{aligned} g(a_i, a_j, r_i, r_j) &= R_{ij}(a_i - a_j) \\ &= \delta_{ij} w(v_i, v_j)(a_i - a_j) \ , \end{aligned} \tag{9}$$

where the δ_{ij} function is 1 if the regions are neighbors and 0 otherwise. This function operates as desired: when a_i is larger than a_j, R_{ij} is greater than 0 and contributes positively to the difference between a_i and a_j. When a_i is smaller than a_j, R_{ij} should be smaller than 0, and contribute negatively to the difference between l_i and l_j. These are modulated by the classifier prediction $w(v_i, v_j)$.

3.5. Learning and Inference

Given the training dataset $T_r = \{(\mathcal{V}_s, \mathcal{R}_s, A_s)\}_{s=1}^{t}\}$ with t samples, where each A_s is the actionness map, we estimate the parameters $\theta = (\alpha, \beta)$ by maximum likelihood. Concretely, the conditional log likelihood of the data is

$$\begin{aligned} L(\theta|T_r) &= \sum_s \log M(A_s|\mathcal{V}_s, \mathcal{R}_s, \alpha, \beta) \\ &= \sum_s \Bigg[\sum_i \alpha f(a_i^{(s)}, r_i^{(s)}) + \\ &\qquad \sum_{i,j} \beta g(a_i^{(s)}, a_j^{(s)}, r_i^{(s)}, r_j^{(s)}) - \sum_i \log Z[\mathcal{R}_s] \Bigg] \end{aligned} \tag{10}$$

where we use the (s) superscript to denote training sample s. We seek the parameter $\hat{\theta}$ that can maximize this log likelihood function. The key to the solution is to integrate $Z(P)$ and then use gradient descent to generate the iteration rules to compute $\hat{\theta}$. By transforming Z to the quadratic formula,

we get (dropping the subscript s on $\mathcal{R}$ for clarity)

$$Z[\mathcal{R}] = \int_z \left(-\alpha z_i^2 + D^T z_i + E\right) \mathrm{d}z \ , \tag{11}$$

$$D = 2\alpha a_i + \beta \left(\sum_j R_{ij} - \sum_i R_{ij} \right) \ ,$$

$$E = -\alpha a_i^2 \ .$$

Based on the properties of the Gaussian distribution, the integration result is

$$Z[\mathcal{R}] = \left(\frac{\alpha}{\pi}\right)^{\frac{t}{2}} \exp\left(\frac{1}{4\alpha} D^T D - \sum_i \alpha a_i^2\right). \tag{12}$$

We then use the gradient descent algorithm to maximize the log likelihood. By maximizing $L(\theta|T_r)$ with respect to $\log\alpha$ and β, the problem is transformed to an unconstrained optimization problem, allowing the direct application of gradient descent. The derivative of $L(\theta|T_r)$ with respect to logα and β are as follows:

$$\frac{\partial L(\theta)}{\partial \log\alpha} = \alpha \sum_s \left[\sum_i -(a_i^{(s)} - \hat{a}_i^{(s)})^2 - \frac{\partial \log Z[\mathcal{R}_s]}{\partial \alpha} \right] \tag{13a}$$

$$\frac{\partial L(\theta)}{\partial \beta} = \sum_s^{n_t} \left[\sum_{ij} R_{i,j}^{(s)} (a_i^{(s)} - a_j^{(s)}) - \frac{\partial \log Z[\mathcal{R}_s]}{\partial \beta}) \right] \tag{13b}$$

The partial derivative $\frac{\partial \log Z[\mathcal{R}_s]}{\partial \alpha}$ and $\frac{\partial \log Z[\mathcal{R}_s]}{\partial \beta}$ are

$$\frac{\partial \log Z[\mathcal{R}_s]}{\partial \alpha} = \frac{t}{2\alpha} - \frac{D^T D}{4\alpha^2} + \frac{D^T a_i}{\alpha} - \sum_i a_i^2 \tag{14a}$$

$$\frac{\partial \log Z[\mathcal{R}_s]}{\partial \beta} = \frac{D^T (\sum_j R_{ij}^{(s)} - \sum_i R_{ij}^{(s)})}{2\alpha} \tag{14b}$$

We incorporate these derivations into the gradient descent algorithm to compute α and β according to Algorithm 1.

Inference Inference on our lattice conditional ordinal random field is straightforward. Since it is a continuous model, we apply the learned parameters and input the test data $(\mathcal{V}_e, \mathcal{R}_e)$ into our model, for a direct solution:

$$\hat{A}_e = \arg\max_{A_e} M(A_e|\mathcal{V}_e, \mathcal{R}_e, \alpha, \beta) \ . \tag{15}$$

We can take the derivative of Eq. 15, set it equal to zero and derive a closed form solution. Each region's actionness is then

$$\hat{a}_i^{(e)} = \frac{2h(r_i^{(e)})\alpha + \beta \left(\sum_j R_{ij}^{(e)} - \sum_i R_{ij}^{(e)} \right)}{\alpha} \ . \tag{16}$$

Algorithm 1: Learning Algorithm of L-CORF

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

3.6. Related Work in Linear Ordering

The linear ordering problem is an NP-hard combinatorial optimization problem with a number of applications such as archaeological seriation and aggregation of individual preferences[13]. Based on the relation between objects to be ranked, Cao et al. [3] proposes a ranking model for the ordering problem in document retrieval setting. The ranking SVM [14] proposes an svm-based ranking method. Both of these two papers rely on local information only for ranking. Kim and Pavlovic [16, 30] introduce a conditional ordinal random field model for dynamic facial emotion prediction and temporal segmentation. Unlike our lattice conditional ordinal random field model, their method only works on the chain-based graphical structure, e.g. temporal segmentation. Qin Tao et al. [26] also propose a continuous Ranking CRF model. The motivation of the model is different from ours and our binary term is more general.

4. Experiments

Data and Features We implement and test our method L-CORF for actionness on both images and videos. For the images, we use Stanford 40 Actions [40], and for videos, we use UCF Sports [29] and Hollywood1 Human Action (HOHA) datasets [22]. Actionness is a new problem; all of these datasets were previously used for action recognition, but they include action bounding boxes and this is what we use for actionness.

The Stanford 40 Action Dataset contains 9532 images of humans performing 40 diverse daily actions, such as riding a bike, playing with guitar and so on. In each image, a bounding box of the person performing the action is provided. All these images come from web resources. The UCF Sports dataset contains 150 videos from 10 action classes, such as diving, golf swinging, walking and so on. The videos are taken from sports broadcasts. The bounding boxes of actions are provided in [38]. HOHA dataset

Table 1: Quantitative comparisons against baselines (mAP).

	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

includes 430 videos with 8 actions, such as answer phone, get out of a car and so on. This dataset is very challenging; significant camera motion, rapid scene changes and background clutter are very common in the videos. Many actions are performed by multiple agents and involve the interactions of them. The bounding boxes[1] of actions in 392 videos are provided by [27]. In these videos, the clips with interesting agents are selected to train and test all the methods.

For computing features, we use basic histograms of oriented gradients (HOG) [4]. On video, we apply the HOG frame-by-frame and the sum and the difference of HOG features are used to represent each cuboid. We select only these features to allow for a fair comparison between our method and baselines, and to emphasize the power of the ordinal random field. Our results show that we achieve a greater of improvement of the proposed models better than the strong baseline of ranking SVM [14], which was used in the objectness paper [7] (see below for a discussion).

Evaluation Protocol In order to evaluate the ranking performance of different methods, we select the mean average precision (mAP) 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 (ie, if a patch overlaps the groundtruth by more than 0.5 then it is scored as positive). Then, PR curves are generated: a recall of k selects the top k ranked patches / cuboids. For these k patches, we compute precision. 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.

We follow the protocol defined by Stanford 40 dataset to assign the training and test examples. The splits for UCF sports and HOHA datasets follows the previous work [20, 22]. In these datasets, we do not distinguish the categories of actions, all the actions are considered as positive samples, non-actions are considered as a negative 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.

4.1. Comparisons with Baselines

Table 1 shows the quantitative comparisons of our L-CORF method against baselines methods. This is the first paper on actionness, so our quantitative comparisons are against relevant baseline methods that could have been used in place of pieces of our method. We use the ranking SVM [14] as a baseline since it was used in a similar visual ranking problem (objectness) [7]. The ranking SVM used the same features as our L-CORF method for this comparison. In both the images and videos, there is a 15+% improvement in our method. For an additional baseline on the video, we apply the moving background subtraction (MBS) method from Shiekh et al. [32], which does not seek to differentiate between general motion and action at all. As we would expect it is unable to perform as well as our method, since intentional motion does not equate to general motion. But it does perform better than the ranking SVM method. This result is also an indicator of the important distinction between motion and action. DPM is another important baseline for both images and videos. It is the state of the art human detector and can be viewed as a method to find actionness by detecting agents. It achieves the best performance on Stanford 40 dataset, so agent detection is useful for actioness detection, although Stanford 40 has limited pose variability. However, actionness detection is quite different from human detection. It does not perform as good as our method on UCF Sports and HOHA datasets.

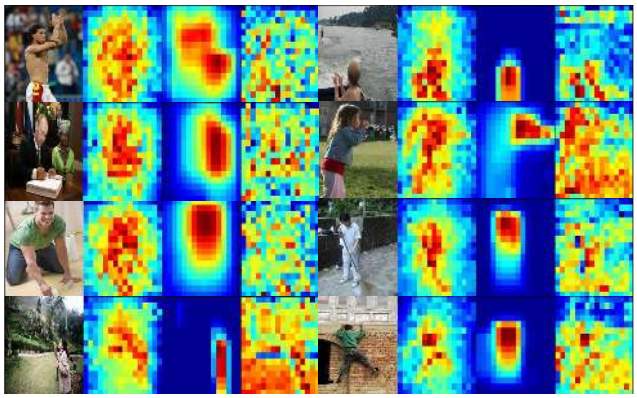
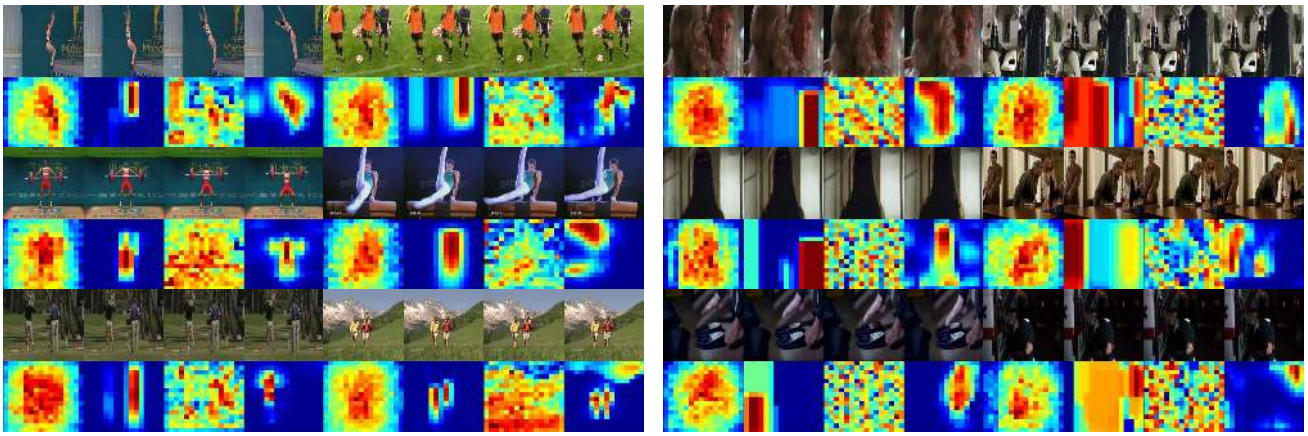


Figure 2: Visual examples of actionness on images from Stanford 40. There are 8 examples (4×2). For each example, the left to right columns are original image, results of L-CORF, DPM and Ranking SVM. DPM is able to effectively detect the human in the image. However, L-CORF is good at finding where where the action happens. The bottom left image is not good result of our method.

We show visual comparisons of our method for both image and video datasets in Figures 2 and 3. We have selected both good and bad cases for our method to present it fairly. In these examples, DPM successfully locates the positions of human, especially for the upright pedestrians, however, some of these persons are not the ones doing the right actions. MBS is able to find the place where the motion is

[1] http://vision.ucla.edu/~raptis/action_part/hoha1_annotations.tar



(a) UCF Sports (b) HOHA

Figure 3: Visual examples of actionness on videos from UCF Sports and HOHA. There are 8 examples (4×2) for each dataset. The first row images are 4 sequential frames within the same cuboid. In the second row, the images are results of L-CORF, DPM, Ranking SVM and MBS from the left to right columns. In these images, we can find that DPM can accurately locate the person, but without considering who is doing the action. MBS is able to accurately detect the motion, but without considering intentional motion. The bottom left images are not good results of our method.

intense. However, general motion is far away from intentional motion (actionness). The bottom left images of all the datasets are bad examples of our method. The one in HOHA dataset is very interesting. The movement of the feet leads to standing up action. Agent body detection is more appropriate here than agent detection. Although DPM performs better than MBS in all the datasets, MBS performs well in this case.

4.2. Performance of Unary Term

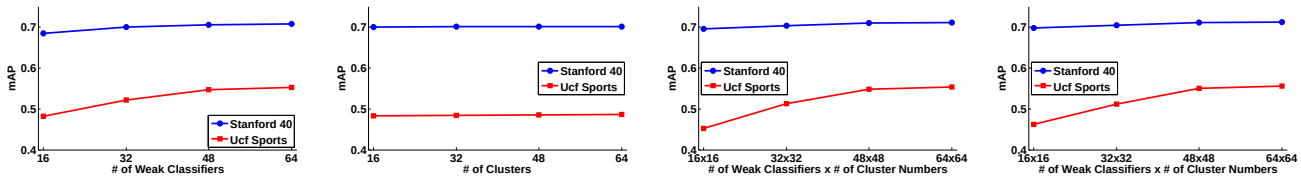
We analyze the unary term by exploring the impact of unary AdaBoost classifier and Hough voting. The number of weak classifiers and the number of the clusters are key parameters for these two methods. We show the performance on both images and video. Figures 4(a) – 4(c) plot the mAP for these two methods seperately and their combination with variant parameters. The variation across the parameter settings is small. From Figure 4(a) and 4(b), we can find that both classifier and hough voting works well for actioness detection on Stanford 40 and UCF Sports datasets. The performance of unary classifier improves with more number of weak classifiers. The performance of unary Hough voting is stable, since it increases slightly with more clusters. The actionness detection on UCF Sports is more difficut than on Stanford 40 dataset, since its mAP is lower. We believe that the actions in UCF Sports have a high variablity than in Stanford 40. Figures 4(a) shows the mAP of only the unary term. The combination of these two decreases the whole performance, comparing to figures 4(a) and 4(b). But when increasing these two numbers, the performance of fusion improves. It is possible that both unary AdaBoost classifier and Hough voting have different advantages.

4.3. Performance of Pairwise Term

An AdaBoost classifier is used to determine the local ranking preference between neighboring regions based on their features. In this experiment, we study the contribution from the binary term for the whole CRF model. We fix the binary AdaBoost classifier with 8 weak classifiers and integrate it into our CRF model. Figure 4(d) demonstrates that the binary term is helpful for the whole CRF model. The improvement is more significant when the unary term has small numbers of weak classifiers and codebook size.

5. Conclusion

Our paper builds on the marked progress in action understanding that has occurred over the last decade. Although this promising work has led to important new methods, our community has not yet studied the interplay between general motion and action. In this paper, we ask exactly that question, define a new notion of actionness and then propose an appropriate ordinal random field model. Our new model incorporates not only local evidence to score a given region's actionness but also takes a rich spatially keyed approach to pairwise order agreement. We have implemented the model on both image and video datasets and achieve strong performance. Our work is the first in this direction and we expect our paper to pave the way for new works on class-independent action analysis and video parsing. In the future, we plan to study the impact of actionness for action



(a) Unary Classifier (b) Unary Hough Voting (c) Unary Term (d) Pairwise Term

Figure 4: The performance of each element in our method. 4(a) shows mAP of the only AdaBoost classifier in the unary term with different numbers of weak classifiers. 4(b) shows mAP of the only Hough Voting in the unary term with different numbers of codebook. 4(c) shows mAP of only the unary term as we variate codebook size and weak classifier numbers. 4(d) shows the performance of our method with binary classifiers, the number of weak classifiers for binary term sets to 8.

detection and recognition tasks. Code for our method and all experiments is available from the author's website.

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

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object ? In *CVPR*, 2010.

[2] D. H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[3] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136. ACM, 2007.

[4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[5] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, 2013.

[6] D. Davidson. Actions, reasons and causes (1963). In *Essays on Actions and Events*. Clarendon Press, Oxford, 2001.

[7] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, pages 575–588, 2010.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010.

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Graph-Based Image Segmentation. *IJCV*, 59(2):167–181, 2004.

[11] R. B. Girshick, P. F. Felzenszwalb, and D. Mcallester. Object detection with grammar models. In *In NIPS*, 2011.

[12] H. Gong and S.-C. Zhu. Intrackability: Characterizing video statistics and pursuing video representations. *IJCV*, 97(3):255–275, 2012.

[13] M. Grötschel, M. Jünger, and G. Reinelt. A cutting plane algorithm for the linear ordering problem. *Operations Research*, 32(6):1195–1220, 1984.

[14] T. Joachims. Making large-scale svm learning practical. *Advances in Kernel Methods - Support Vector Learning*, 1999.

[15] G. Johanssson. Visual-perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 1973.

[16] M. Kim and V. Pavlovic. Structured output ordinal regression for dynamic facial emotion intensity prediction. In *ECCV*, 2010.

[17] O. Klliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *ECCV*, 2012.

[18] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, 2013.

[19] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011.

[20] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011.

[21] I. Laptev. On space-time interest points. *IJCV*, 2005.

[22] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[23] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009.

[24] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104:90–126, 2006.

[25] K. A. Pelphrey, T. V. Mitchell, M. J. McKeown, J. Goldstein, T. Allison, and G. McCarthy. Brain activity evoked by the perception of human walking: Controlling for meaningful coherent motion. *The Journal of Neuroscience*, 23(17):6819–6825, 2003.

[26] T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang, and H. Li. Global ranking using continuous conditional random fields. In *NIPS*, 2008.

[27] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012.

[28] H. Ren and G. Xu. Human action recognition in smart classrooms. In *FG*, 2002.

[29] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[30] O. Rudovic, V. Pavlovic, and M. Pantic. Kernel conditional ordinal random fields for temporal segmentation of facial action units. In *ECCV Workshops*, 2012.

[31] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[32] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. In *ICCV*, 2009.

[33] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *CVPR*, 2004.

[34] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[35] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[36] T. Wang, X. He, and N. Barnes. Learning structured hough voting for joint object detection and occlusion reasoning. In *CVPR*, 2013.

[37] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *CVPR*, 2012.

[38] R. Xu, P. Agarwal, S. Kumar, V. N. Krovi, and J. Corso. Combining skeletal pose with local motion for human activity recognition. In *AMDO*, 2012.

[39] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010.

[40] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.