Discriminative models for static human-object interactions

Chaitanya Desai
Univ. of California, Irvine
desaic@ics.uci.edu

Deva Ramanan
Univ. of California, Irvine
dramanan@ics.uci.edu

Charless Fowlkes
Univ. of California, Irvine
fowlkes@ics.uci.edu

Abstract

We advocate an approach to activity recognition based on modeling contextual interactions between postured human bodies and nearby objects. We focus on the difficult task of recognizing actions from static images and formulate the problem as a latent structured labeling problem. We develop a unified, discriminative model for such context-based action recognition building on recent techniques for learning large-scale discriminative models. The resulting contextual models learned by our system outperform previously published results on a database of sports actions.

1. Introduction

One of the long standing goals in computer vision is to build a computer system that can see people and recognize their activities. There is general agreement that taxonomies of activities should be hierarchical [2, 13] - complex activities such as "eating lunch" are composed out of simpler low-level actions such as "picking up a fork". Many low-level actions that form the foundation for such a hierarchy involve interactions between humans and objects. This work examines the task of recognizing such interactions from static images through the use of discriminative models.

One of our motivations for representing actions as contextual human/object interactions comes from the medical literature on nursing and motor rehabilitation [21, 5]. Standard clinical assessments of motor ability typically define a set of everyday actions required to perform *activities of daily living*, such as picking up a telephone, drinking from a mug, and turning on a light switch. Such taxonomies suggest that actions are not performed in isolation, but are typically done with the goal of manipulating nearby objects.

Contextual models for high-level reasoning: There has been much work on contextual reasoning for object recognition, typically focused on the task of object detection. While [25, 28, 18] convincingly demonstrate that context can refine the output of weak local detectors, such improvement is harder to show for highly-tuned detectors [11, 7, 8]. In the extreme case, a perfect detector will exhibit no improvement due to contextual reasoning. Even given such perfect detectors, we argue that context is *still* necessary for higher-level inferences such as action recognition. Identifying a "drinking" action is more subtle than simply finding a bottle because one must verify that a bottle and person satisfy particular spatial relationships and poses.

Our contributions: In this paper, we demonstrate that context *does* provide a strong improvement over the same state-of-the-art local detectors in [11, 7, 8] when evaluated for the task of action recognition. Furthermore, while past work on object-object context model relative locations of objects [14, 7, 28], we demonstrate that *pose* is a crucial component of these relations - a tennis serve is defined not simply by the location of a person and racket, but also by their poses. Finally, we introduce a unified, discriminative action model for simultaneously learning object templates,

Figure 1. On the **left**, we show examples of everyday actions from the PASCAL VOC dataset [10] that involve interactions between postured bodies and objects. We develop discriminative action models that capture these interactions. To demonstrate our approach using standard action recognition benchmarks, we evaluate our models on action categories from the database of [16] (with example images shown on the **right**). This database of sports actions contains images of human bodies interacting with objects. We show that action models that capture spatial interactions between bodies and objects outperform state-of-the-art local detectors on this dataset.

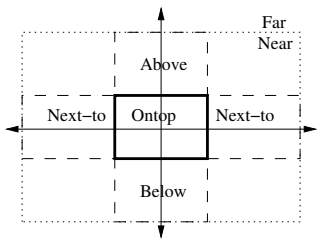

Figure 2. A visualization of our implementation of the spatial histogram feature d_{ij} from [7]. We consider the location of the center of window j with respect to a coordinate frame defined by window i, denoted by the thickly outlined box. The dashed and dotted rectangles represent regions over which the center of window j are binned. The relative location of j must either be `far` or `near`. For near windows, we consider `above`, `ontop`, `below`, and symmetric `next-to` bins as shown. This makes d_{ij} a 6 dimensional sparse binary vector.

pose-based templates, and their contextual relations.

2. Related work

Actions: The literature on action recognition is too large to review here - we refer the reader to the recent survey [13]. Many approaches are based on local motion features such as periodicity [26, 6], spatio-temporal gradients [37], optical flow [9, 27], or background-subtracted silhouette features [3, 15]. Local scores from such features are often integrated across time using a Hidden Markov Model (HMM) [33, 4].

Context: Our contextual approach to action recognition is heavily motivated by the work of Gupta, Kembhavi, and Davis [16] and Laptev and Perez [23]. Both works focus on contextual analysis of actions in the video-domain but [16] also present a static-image version of their model. In concurrent work, Yao and Fei-Fei also examine the use of contextual models of human-object interactions [35, 34]. In particular, [35] introduce a discriminative contextual model similar to ours, but defined on articulated poses rather than discrete mixtures-of-poses. Context in image recognition has been an active area of research in recent history [28, 25, 14, 17, 18, 1, 22, 7]. These approaches have typically treated the problem as that of finding a joint labeling for a set of pixels, super-pixels, or image segments and are usually formulated as a CRF.

Still images: Wang et al [31] provide one of the earliest examples of static-image action recognition in computer vision. Ikizler et al [19] present a similar approach to recognizing actions based on human pose estimation. Gupta et al [16] examine contextual cues for static-image action recognition using generative probabilistic models.

3. Model

We propose a unified model for action recognition from static images based on detecting spatial conjunctions of multiple objects. Our formulation utilizes the work of [7] as an internal representation of the contextual relations between objects. Although we utilize a structured representation very similar to that in [7], the output of our system is a K-way classification (action category) rather than a structured output (collection of object detections). We thus treat object detections with their relative pose and spatial locations as a set of latent variables which we maximize over at test time.

We now review the contextual representation from [7] as it forms the basis of our action framework.

Image model: Our representation of an image is a collection of M overlapping windows at various locations. The location of the i^{th} window is given by its center and scale, written as $l_i = (x, y, s)$. The orientation of some manipulable objects may be important to model, in which case one can augment l_i to also include orientation θ. Let x_i denote the features extracted from the i^{th} window. We use a histogram of oriented gradients (HOG) descriptor. The entire image will be represented as a collection of feature vectors $X = \{x_i : i = 1 \dots M\}$.

Action model: Let $S_y(X)$ be the score associated with a particular action label for image X, where y takes on one of K action classes:

$$y^* = \underset{y}{\operatorname{argmax}}\, S_y(X) \quad \text{where} \quad y \in \{0 \dots K\} \tag{1}$$

The score associated with particular action labeling for an image is obtained by searching over all possible configurations of objects consistent with that action model. We denote the configuration of objects in an image as Z:

$$S_y(X) = \max_Z S_y(X, Z) \tag{2}$$

To define Z, we suppose that there exist O different object models spread over the K actions of interest. We write $z_i \in \{0 \dots O\}$ for the label of the i^{th} window, where the 0 label designates background. Then $Z = \{z_i : i = 1 \dots M\}$ is a vector of labels that identifies which windows in an image contain particular objects.

Contextual model: We define the score of labeling image X with object configuration Z given a particular action y as:

$$S_y(X, Z) = \sum_{i,j} w^y_{z_i, z_j} \cdot d_{ij} + \sum_i w_{z_i} \cdot x_i \tag{3}$$

Here, $w^y_{z_i, z_j}$ represent weights that encode valid geometric configurations of object classes z_i and z_j for action y, whereas w_{z_i} represents a local template for object class i. As in [7], d_{ij} is a spatial context feature that bins the relative location of windows i and j into one of D canonical relations including above, below, overlapping, next-to, near, and far (Fig2). Hence d_{ij} is a sparse binary vector of length D with a 1 for the kth element when the kth relation

is satisfied between the current pair of windows. Note that the contextual model is action-specific, but the local object templates are not. For example, tennis rackets and tennis balls both appear in various tennis actions, but their spatial relationships differ for a "forehand" versus a "serve" action.

Pose model: Intuitively, the pose of the human body plays a large role in defining a body-object interaction. One simple mechanism for capturing pose is a mixture model. In this case, we define a local template for each mixture component $w_{z_i,\theta}$ where $\theta \in \{0 \dots T\}$. We would like to learn a separate pairwise model w_{z_i,z_j} conditioned on the mixture component for each object. For example, we expect manipulatable objects to be positioned near the outstretched hand of a person during interactions. We can formalize this in a straightforward manner by augmenting the label space to be the cross product of object class labels and mixture components:

$$z_i \in \{0 \dots O\} \times \{1 \dots T\} \tag{4}$$

Given the above extension, we can use the pairwise model from (3) to capture pose-specific contextual models. From the point-of-view of the contextual model, different mixture components can be thought of as different objects. As such, we henceforth write $\hat{O} = OT$ for the number of distinct object-mixture combinations. We show in our experimental results that the extension to pose-specific mixtures, though somewhat trivial in terms of notation, is crucial for building accurate action models. This is because human body posture is a strong cue for action recognition.

Sparsity: As in [7], we constrain local and pairwise background weights w_0 and $w^y_{i,0}$ and $w^y_{0,i}$ to be 0. Since the majority of windows in an image will be labelled as background, this significantly speeds up computations with the model.

Action-specific object vocabularies: Only a subset of objects take part in a particular action - one does not typically encounter volleyballs in tennis. Similar, only a subset of poses take part in a particular action - humans do not stand in a bowling pose while playing croquet. We can enforce such constraints by fixing $w^y_{i,j} = -\infty$ unless both object-mixture component i and component j occur in at least one instance of action y in the training data. Though our learning algorithm can naturally learn such constraints, we found fixing them to be easier since this (1) simplifies inference since ony a small number of objects need to be considered when scoring an action (Sec.4) and (2) significantly reduces the number of pairwise parameters that need to be learned (Sec.5).

Single-instances: In general, images may contain multiple instances of objects. We will consider a restricted version of the action recognition problem where at most a single instance of each object is present (e.g., each image has at most one person). While not crucial to action classification performance, we found this restriction to greatly improve quality of detected objects across current action datasets. In our framework, this restriction can easily be enforced by setting $w^y_{i,i} = -\infty$ for all action labels y and objects classes i.

4. Inference

Given a collection of detection windows X for an image, we want to compute:

$$y^* = \operatorname*{argmax}_y \max_Z S_y(X, Z) \tag{5}$$

This requires finding, for each action, the best-scoring configuration of objects for that image. The computational bottleneck is the inner maximization over window labels Z which is intractable for general pairwise potentials. One may resort to search techniques such as branch-and-bound or A* to find the exact maximum, but we find that a simple greedy forward search is usually sufficient.

Greedy forward selection: We adopt a the procedure described in [7] which is inspired be greedy algorithms traditionally used for non-maximum suppression [24]. First, we initialize Z by labeling each window as background $z_i = 0$. We then repeat the following:

1. Compute the single window and object-mixture label z^*_i that increases the score $S_y(X, Z)$ by the largest amount.
2. Update Z with z^*_i.

The above is repeated until instancing any other window decreases the total score. Naively recomputing the scores associated with instancing all possible remaining windows takes prohibitively long, but one can incrementally keep track of the potential gain of turning on a window as described in [7]. We have found that in practice this greedy algorithm performs as well as other approximation techniques (e.g., tree-reweighted belief propagation) and runs very fast.

5. Learning

Max-margin learning: Suppose we are given triples of $\{X_n, Z_n, y_n\}$ collection of training images, object configurations, and action labels respectively. We want to train a model w that given a new image X_n, tends to produce the true action label. We formulate this as a regularized leaning problem:

$$\arg \min_{w, \xi_n \geq 0} \frac{1}{2} w^T w + C \sum_n \xi_n \tag{6}$$

$$\text{s.t.} \forall n, y \neq y_n \quad S_{y_n}(X_n, Z_n) - S_y(X_n) \geq l(y_n, y) - \xi_i$$

where we recall that $S_y(X_n) = \max_Z S_y(X_n, Z)$. The constrained optimization from (6) has the following interpretation: Consider the n^{th} training image. We want the score of the true action model y_n and object configuration Z_n to dominate the score of any alternative action model y, over any object configuration Z. However, not all alternative actions are equally bad. One may wish to pay a higher penalty for mislabeling a "tennis-forehand" as a "volleyball-smash" rather than a "tennis-serve". The loss function $l(y_n, y)$ measures the cost for mislabeling action y_n as y, and penalizes the slack variable ξ_n in proportion. This manner of loss augmentation is known as margin-rescaling [30].

We use a simple 0-1 loss:

$$l(y_n, y) = \mathbf{1}_{[y_n \neq y]}$$

where $\mathbf{1}$ is the standard identify function. Note that since we are performing classification rather than structured prediction, there is no requirement that the loss function decompose over detection windows (as in the model of [7]).

Linear score: Our method for learning the parameters uses the fact that the score in (5) can be written as a single inner product of a weight vector and feature vector. Let us re-write the scoring function from (3) by introducing unary as well as pairwise potential functions as follows:

$$S_y(X, Z) = \sum_{i,j} w_s^y \cdot \psi(z_i, z_j, d_{ij}) + \sum_i w_a \cdot \phi(x_i, z_i)$$

where w_s^y are the weights on the pairwise potentials for action y and w_a are the weights on the unary potentials. The length of vectors w_s^y and $\psi()$ is $D\hat{O}^2$, and w_a and $\phi()$ are vectors of length $\hat{O}F$, where D is the number of spatial relations, $\hat{O}$ is the number of object-mixture combinations, and F is the length of feature vector x_i. The vector $\psi()$ will contain at most D nonzero entries and the vector $\phi()$ will contain F nonzero entries. We can now write the score for object configuration Z under action y as an inner product:

$$S_y(X, Z) = w \cdot \Psi(X, Z, y) \quad \text{where} \tag{7}$$

$$w = \begin{bmatrix} w_s^1 \\ w_s^2 \\ \vdots \\ w_s^K \\ w_a \end{bmatrix} \quad \text{and} \quad \Psi(X, Z, y) = \begin{bmatrix} \vdots \\ 0 \\ \sum_{ij} \psi(z_i, z_j, d_{ij}) \\ 0 \\ \vdots \\ \sum_i \phi(x_i, z_i) \end{bmatrix}$$

is a long vector with zeros everywhere except for the "slot" corresponding to the y^{th} action, which will contain the pairwise potentials corresponding to the y^{th} action, and the last element which will contain the unary potentials.

Structural SVM: The constraints from (6) are nonlinear in the parameters w (since they contain an embedded max over Z). To make them conceptually easier to work with, we can rewrite each single constraint as a set of linear constraints by enumerating all possible Z's from the max over object configurations:

$$\arg \min_{w, \xi_n \geq 0} \frac{1}{2} w^T w + C \sum_n \xi_n \tag{8}$$
$$\text{s.t.} \forall n, Z, y \neq y_n, \quad w \cdot [\Psi(X_n, Z_n, y_n) - \Psi(X_n, Z, y)] \geq l(y_n, y) - \xi_i$$

Note that (8) and (6) represent equivalent problems. The training problem specified in (8) is a quadratic program with an exponential number of linear constraints known as a structural SVM [29].

Cutting-plane optimization: Even though (8) has an exponential number of constraints, there tends to be a small number of active constraints at the optimum solution – the "support vectors". One can use the excellent public package SVMStruct [20] to compute such solutions. We found it more convenient to implement our own cutting plane solver. The computational bottleneck of the optimization is the step that computes the most violated constraint for an image X_n:

$$(y^*, Z^*) = \operatorname*{argmax}_{y \neq y_n, Z} l(y_n, y) + w \cdot \Psi(X_n, Z, y) \tag{9}$$

The above argmax can be computed as follows: For each action class y other than the true action y_n, we compute the score of the best object configuration $\max_Z w \cdot \Psi(X_n, Z, y)$ using the greedy procedure from Sec.4, add the cost $l(y_n, y)$, and finally return the class y^* and object configuration Z^* that produce the maximum score. The pair (y^*, Z^*) represents the false action, configuration combination that is most likley to be selected by the action classifier – in structural SVM terms, the most-violated constraint for image X_n given the current model w.

Because greedy algorithms are an under-generating approximation [12], we loose formal guarantees of optimality at convergence of the cutting plane algorithm outlined in [20]. However, we observe that in practice the greedy forward search tends to produce scores similar to brute-force solutions, suggesting that we may still learn models that are close to optimal in practice.

Latent Structural SVMs: Because we are scoring ourselves in multi-way classification, the action configurations $\{Z_n\}$ are auxiliary labels that do not directly affect our loss function. Put in other words, one can treat the ground-truth action configuration $\{Z_n\}$ as latent variables that can also be estimated during learning so as to minimize the overall loss. Such a model is known as a latent structural SVM [36] or a max-margin hidden conditional random field [32]. Because the formulation is no longer convex, the training

procedure is sensitive to initialization. However, it is natural to initialize action configurations to the given labels $\{Z_n\}$. Our experiments so far have suggested that such a procedure results in overfitting, as training error always decreased but testing error increased. We see this avenue as an important direction for future exploration.

6. Experimental results

It appears difficult to build large, realistic datasets of interesting actions. Most video action datasets are relatively contrived, consisting of actors performing scripted actions on simplistic backgrounds. This is one of our motivations for exploring image-based action representations rather than video-based representations. We only know of one publically available image-based action dataset used in the recent work of [16]. This dataset contains images of humans engaged in 6 different sports actions, defined by a variety of body postures and contextual object relations. Although this dataset is small by contemporary standards (50 images per action, split into a train and test set), it provides a good starting point for this new problem. [16] describe a fairly complex graphical model defined on local HOG detectors and segmentation masks, which is trained on human silhouettes and additional images not included in the benchmark. We will show that our simpler, unified model achieves significantly better performance using only the given benchmark training data.

Training: Rather than learning the local templates in our framework, we used the scores output by the Felzenwzwalb et al. [11] detector as a single feature. Recent benchmarks [10] indicate this system is the current state-of-the-art object detection system. To learn biases between different object classes, we append a constant 1 to make x_i two-dimensional.

Results: Figure 5 presents the performance of our system on the benchmark dataset of [16]. Our contextual action model achieves an average classification score of 82.5%, outperforming the best reported score of 78.7%. We also compare our system to a simple "bag of objects" action model built using appropriately bias-aligned local detectors. Such a baseline actually performs reasonably well, beating all of the baselines presented in [16]. Some example results of the activity classification and latent object labels are shown in Figure 6.

Importance of pose: We also compare against a baseline that doesn't include object pose. As shown in Fig.5, action representations of object configurations that ignore body pose perform quite poorly, scoring 42%. This fact suggests that spatial interaction models of objects should not be built from pose-invariant detectors! A tennis-forehand action is not defined purely by the relative location of a racket and human, but also by the orientation of the racket and the posture of the body. While it may be difficult to design representations that factor into account continuous notions of pose, we show that a simple mixture-model captures a considerable amount of information.

The power of context: Our performance results are noteworthy as they demonstrate that contextual models of spatial interaction provide a stark improvement over the state-of-the-art local models from [11] - our average classification rates improve from 70% to 82%.

While our action-specific contextual models utilize the discriminative contextual features of Desai et al. [7] as a latent or auxiliary component, we get far larger gains in performance (they report an improvement from 26% to 27% in average precision). Indeed, many recent benchmark evaluations of context in object detection, such as [8, 14], suggest that there is relatively little to be gained by contextual reasoning when compared to highly-tuned local detectors. Our results indicate that such context is indeed helpful for higher-level inferences beyond object detection.

7. Conclusion

Most existing models of actions examine body posture. We argue that interactions of nearby objects should also be modeled for action recognition, as such contextual information helps uncover the low-level goals and purpose of the performed action. We have developed a simple but accurate discriminative model of human-object interactions, and demonstrated that it clearly outperforms highly-tuned local detectors for the task of action recognition.

Acknowledgments Funding for this research was provided by NSF Grant IIS-0954083 and NSF Grant IIS-0812428.

References

[1] R. Baur, A. A. Efros, and M. Hebert. Statistics of 3d object locations in images. Technical Report CMU-RI-TR-08-43, Robotics Institute, Pittsburgh, PA, October 2008. 2

[2] A. Bobick. Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical Transactions: Biological Sciences*, 352(1358):1257–1265, 1997. 1

[3] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE PAMI*, 23(3):257–267, 2001. 2

[4] M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. In *IEEE CVPR*, pages 994–999, 1997. 2

[5] A. Catz, M. Itzkovich, E. Agranov, H. Ring, and A. Tamir. SCIM-spinal cord independence measure: a new disability scale for patients with spinal cord lesions. *Spinal Cord*, 35(12):850–856, 1997. 1

[6] R. Cutler and L. Davis. Real-time periodic motion detection, analysis, and applications. In *IEEE CVPR*, volume 2, pages 326–332, 1999. 2

[7] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models of multi-class object layout. In *ICCV*, 2009. 1, 2, 3, 4, 5, 7

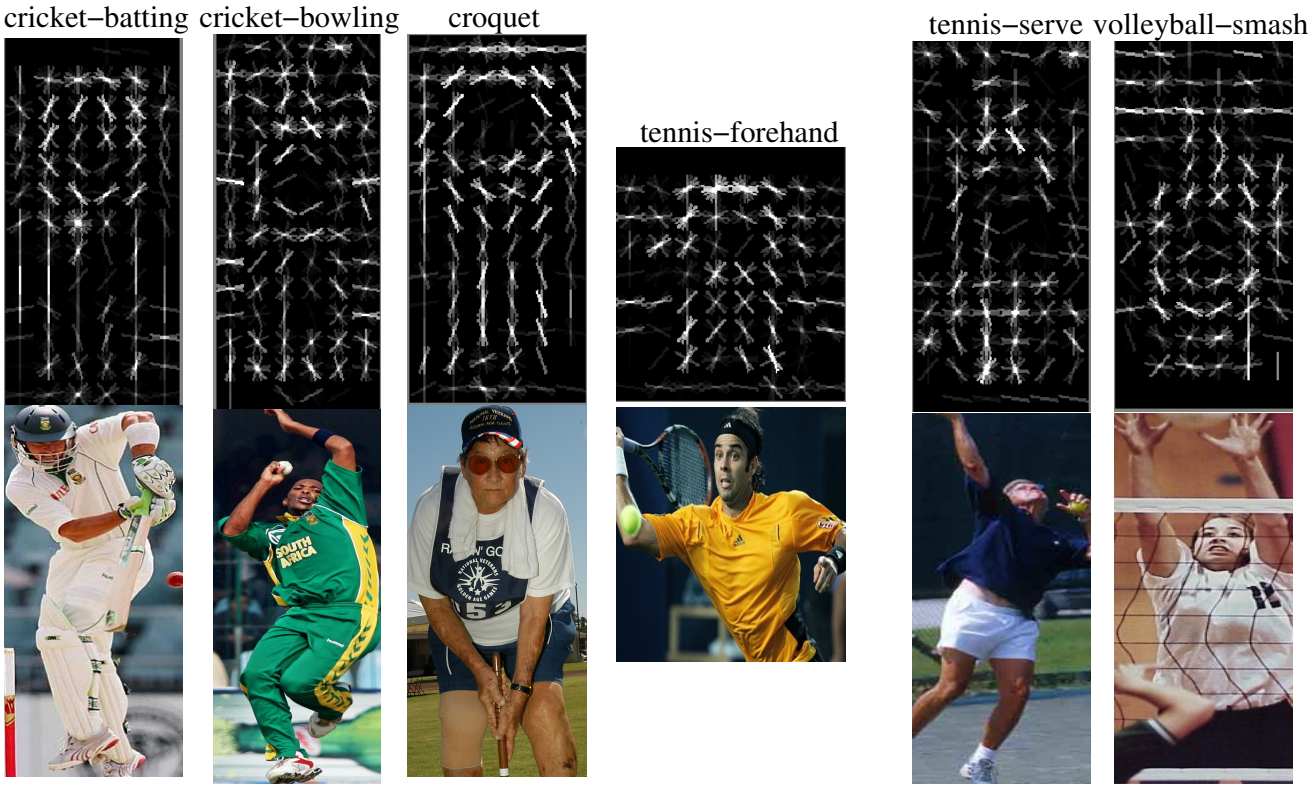

Figure 3. We learn HOG weights for 6 different poses, one associated with each action using the local models from [11]. We show the learned root templates above example training images, omitting the part templates for simplicity. By learning multi-posed local models, we are able to in-turn learn pairwise weights w_{ij}^{y} that are conditioned on these latent poses. We show that pose-aware contextual models are vital for action recognition in Fig.5.

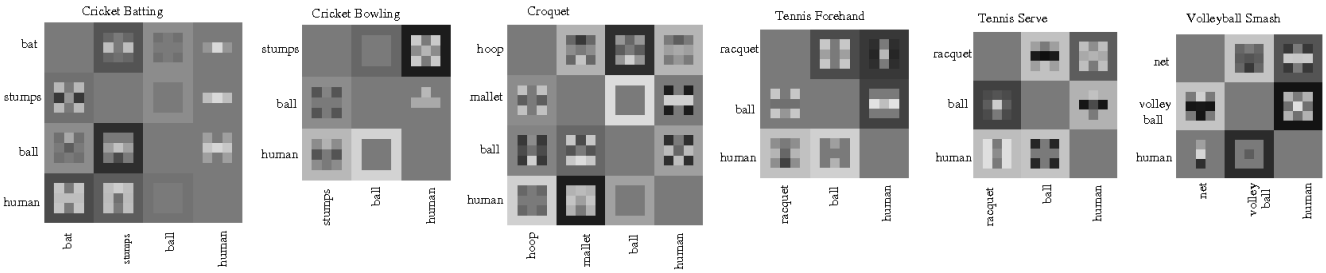

Figure 4. A visualization of the spatial weights across pairs of classes for each of the 6 actions. There are 14 object types spread over the 6 actions. However, only a fraction of these participate for any action (4 at most). The weights for all other pairs of classes are zero. Light areas indicate a favorable arrangement. The weight is for the object in the row w.r.t. the object in the column. For instance, in the case of *cricket batting*, the model favors seeing stumps next to and on top of humans. For *cricket bowling*, the model favors seeing the ball above the human. For *croquet*, the mallet tends to be on top of or next to the human. Likewise, for *volleyball smash*, the model favors seeing volleyballs above the net. The all-zero diagonal entries across actions reflects the fact that we do not learn any pairwise relationships for objects of the same class.

[8] S. Divvala, D. Hoiem, J. Hays, and A. Efros. An empirical study of context in object detection. In *IEEE CVPR*, 2009. 1, 5

[9] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE CVPR*, volume 2, pages 726–733, 2003. 2

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html. 1, 5

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 99(1), 5555. 1, 5, 6

[12] T. Finley and T. Joachims. Training structural SVMs when exact inference is intractable. In *International Conference on Machine Learning (ICML)*, pages 304–311, 2008. 4

[13] D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ra-

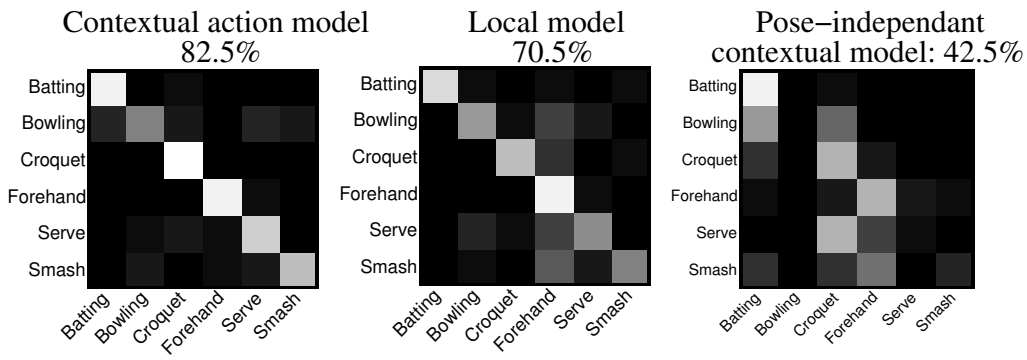

Figure 5. Class-confusion matrices on contextual action dataset from [16]. Our average classification score of 82.5% (**left**) is better than the score of 78.76% reported in [16], but slightly below the score of 83.3% achieved by Yao and Fei-Fei [35]. Our results are impressive given we only use the given bounding-box data, while [35] requires articulated body labels and [16] require human silhouette segmentations as well as additional images of objects to train their local detectors. We also show results for a "bag-of-objects" action model obtained by omitting the pairwise terms from (3). This model achieves a score of 70.5% (**center**), outperforming the scene and pose-based baselines of [16] (scoring 65.83% and 57.5% respectively). Finally, we show an additional baseline (**right**) obtained by a naive implementation of [7] that learns a contextual action model built on a single pose-invariant human detector across all classes. The single detector is still formulated as a mixture model over poses, but the contextual pairwise weights w_{ij}^{y} are not learned separately for each mixture component. It performs significantly worse, indicating that body pose is an important cue for capturing spatial interactions.

Figure 6. We show results of our system. The top 2 rows show correct predictions and the 3rd row shows incorrect predictions. Many of the mistakes can be corrected with improved local detectors. For the 1st and 2nd mistakes, the bowler's hand getting confused with a tennis racquet and the absence of cricket stump detections leads to the human bowler pose being incorrectly detected as a tennis serve pose. For the 3rd misclassification, the volleyball getting confused with a small ball and the absence of a volleyball net detection lead to the volleyball smash action being misclassified as a human bowling action. For the 4th mistake, the tennis racquet firing near the hand and the volleyball not being detected lead to the human volleyball pose being detected as a human forehand pose.

manan. Computational studies of human motion i: Tracking and animation. *Foundations and Trends in Computer Graphics and Vision*, 1(2/3):1–255, 2006. 1, 2

[14] C. Galleguillos, A. Rabinovich, and S. Belongie. Object cat-

egorization using co-occurrence, location and appearance. In *IEEE CVPR*, Anchorage, AK, 2008. 1, 2, 5

[15] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE PAMI*, 29(12):2247, 2007. 2

[16] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE PAMI*, 31(10):1775–1789, 2009. 1, 2, 5, 7

[17] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *IEEE CVPR*, volume 2. IEEE Computer Society; 1999, 2004. 2

[18] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 80(1):3–15, 2008. 1, 2

[19] N. Ikizler, R. Cinbis, S. Pehlivan, and P. Duygulu. Recognizing actions from still images. In *ICPR*, pages 1–4, Dec. 2008. 2

[20] T. Joachims, T. Finley, and C. Yu. Cutting plane training of structural SVMs. *Machine Learning*, 2009. 4

[21] B. Kopp, A. Kunkel, H. Flor, T. Platz, U. Rose, K. Mauritz, K. Gresser, K. McCulloch, and E. Taub. The Arm Motor Ability Test: reliability, validity, and sensitivity to change of an instrument for assessing disabilities in activities of daily living. *Archives of physical medicine and rehabilitation*, 78(6):615–620, 1997. 1

[22] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, volume 2, 2005. 2

[23] I. Laptev and P. Perez. Retrieving actions in movies. In *International Conference on Computer Vision*, 2007. 2

[24] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 17–32, 2004. 3

[25] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: a graphical model relating features, objects and scenes. *NIPS*, 16, 2003. 1, 2

[26] R. Polana and R. Nelson. Detection and recognition of periodic, nonrigid motion. *IJCV*, 23(3):261–282, 1997. 2

[27] E. Shechtman and M. Irani. Space-time behavior based correlation. In *IEEE PAMI*, volume 29, pages 2045–2056, November 2007. 2

[28] A. Torralba, K. Murphy, and W. Freeman. Contextual models for object detection using boosted random fields. *NIPS*, 2004. 1, 2

[29] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*. ACM New York, NY, USA, 2004. 4

[30] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research (JMLR)*, 6:1453–1484, September 2005. 4

[31] Y. Wang, H. Jiang, M. Drew, Z. Li, and G. Mori. Unsupervised discovery of action classes. In *IEEE CVPR*, pages 1654–1661, 2006. 2

[32] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. Technical report, Technical Report TR 2008–21). School of Computing Science, Simon Fraser University, 2008. 4

[33] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *IEEE CVPR*, pages 379–385, 1992. 2

[34] B. Yao and L. Fei-Fei. Grouplet: a structured image representation for recognizing human and object interactions. In *CVPR*. 2

[35] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*. 2, 7

[36] C. Yu and T. Joachims. Learning structural SVMs with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM New York, NY, USA, 2009. 4

[37] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *CVPR*, volume 2, 2001. 2