Classifying Actions and Measuring Action Similarity by Modeling the Mutual Context of Objects and Human Poses

Bangpeng Yao BANGPENG@CS.STANFORD.EDU
Aditya Khosla ADITYA86@CS.STANFORD.EDU
Li Fei-Fei FEIFEILI@CS.STANFORD.EDU
Computer Science Department, Stanford University, Stanford, CA 94305, USA

Abstract

In this paper, we consider two action recognition problems in still images. One is the conventional action classification task where we assign a class label to each action image; the other is a new problem where we measure the similarity between action images. We achieve the goals by using a mutual context model to jointly model the objects and human poses in images of human actions. Experimental results show that our method not only improves action classification accuracy, but also learns a similarity measure that is largely consistent with human perception.

1. Introduction

Human action recognition in still images is attracting much attention in computer vision (Laptev & Mori, 2010). Many recent works use contextual information (Gupta et al., 2009; Yao & Fei-Fei, 2010b) to help improve the recognition performance. Compared to the methods that directly associate low-level image descriptors with class labels (Yao & Fei-Fei, 2010a; Delaitre et al., 2010), context (e.g. estimating human pose, detecting objects) provides deeper understanding of human actions.

Following the method of Yao & Fei-Fei (2010b), in this paper we consider human actions as interactions between humans and objects, and jointly model the relationship between them using the *mutual context model*. As shown in Fig.1, our method allows objects and human poses to serve as mutual context to facilitate the recognition of each other, based on which we address two action recognition tasks:

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

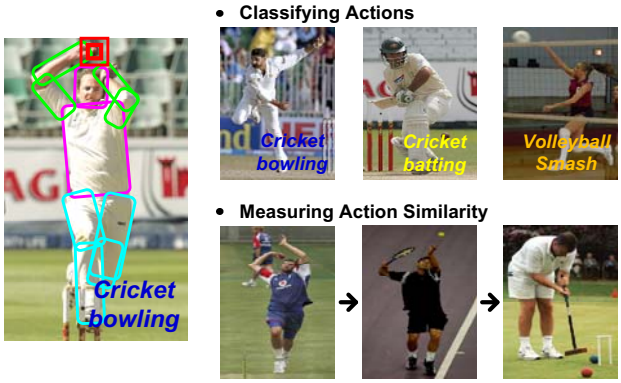
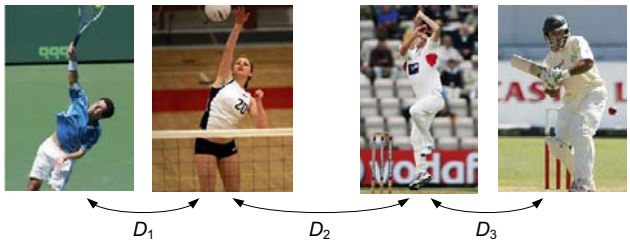


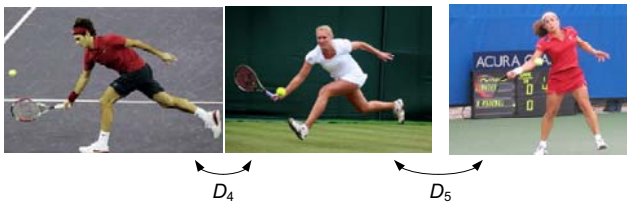
Figure 1. Objects and human poses can facilitate the recognition of each other in the actions of human-object interactions, as shown in the cricket bowling image. Based on the recognition of objects and human poses, we consider two tasks: action classification and measuring action similarity. "→" indicates that the left image is more similar to the left-most cricket bowling image than the right one.

- Conventional *action classification* where we assign a class label to each action image.
- *Measuring the similarity* between different action images. The goal is to make the similarity measure consistent with human perception.

The second task, measuring action similarity, is very different from conventional action classification problems. As shown in Fig.2, human actions lie in a relatively continuous space and different actions can be correlated. We humans are able to distinguish small changes in human poses as well as capture the relationship of different actions from the objects or scene background. However it is difficult to capture all these subtleties by simply assigning action images into several independent classes as in the conventional action classification problem. In this work, by explicitly considering objects and human poses, we obtain a dis-



(a) A human action can be more related to some actions than others. $D_1 < D_2$ because the left-most two images have similar human poses. $D_3 < D_2$ because the right-most two images are from the same sport and the objects "cricket ball" and "cricket stump" are present in both images.



(b) Human actions lie in a continuous space. Humans are able to capture the difference between different images even if they belong to the same action class. $D_4 < D_5$ because the left two images have very similar human poses.

Figure 2. Examples of the distance between different images of human actions denoted by D_i.

tance[1] measure of action images which is largely consistent with human annotation.

1.1. Related Work

Our method builds upon the mutual context model (Yao & Fei-Fei, 2010b) that explores the relationships between objects and human poses in human actions. The model presented in this paper is more flexible and discriminative in that: (1) it learns an overall relationship between different actions, objects, and human poses, rather than modeling each action class separately; (2) it can deal with any number of objects, instead of being limited to the interactions between one human and one object; (3) it incorporates a discriminative action classification component which takes global image information into consideration.

While different objects and annotations of action classes can be represented by discrete indexes, human poses lie in a space where the location of body parts changes continuously. To make the joint modeling of actions, objects, and human poses easier, we discretise possible layouts of human body parts into a set of representative poses, termed as *atomic poses* (as shown in Fig.3). Our atomic poses are discovered in a similar manner as poselets (Bourdev & Malik, 2009). While poselets are local detectors for specific body parts, the atomic poses consider the whole human body and can be thought of as a dictionary of human poses.

In the rest of this paper, we first elaborate on the mutual context model and distance measure method in Sec.2, and then present experimental results in Sec.3.

2. Algorithm

In this section, we describe the mutual context model that jointly models a set of actions $\mathcal{A}$, objects $\mathcal{O}$, and atomic poses $\mathcal{H}$. We first introduce the model (Sec.2.1), then describe how to obtain the atomic poses (Sec.2.2) and the model learning approach (Sec.2.3). Finally we show our approach to classify action images (Sec.2.4) and measure action distance (Sec.2.5).

2.1. Mutual Context Model Representation

Given an image I with annotations of action class $A \in \mathcal{A}$, bounding boxes of objects $O \in \mathcal{O}$ and body parts in the human pose $H \in \mathcal{H}$, our model learns the strength of the interactions between them. We further make the interaction conditioned on image evidence, so that the components that are harder to recognize play less important roles in the interaction. Our model is represented as

$$\begin{aligned}\Psi(A,O,H,I) &= \phi_1(A,O,H) + \phi_2(A,I) \\ &+ \phi_3(O,I) + \phi_4(H,I) + \phi_5(O,H)\end{aligned} \quad (1)$$

where ϕ_1 models the compatibility between A, O, and H; ϕ_{2-4} models the image evidence using state-of-the-art action classification, object detection, and pose estimation approaches; ϕ_5 considers the spatial relationship between objects and body parts. We now elaborate on each term in Eqn.1.

Compatibility between actions, objects, and human poses. $\phi_1(A,O,H)$ is parameterized as

$$\begin{aligned}&\phi_1(A,O,H) \quad (2)\\ &= \sum_{i=1}^{N_h}\sum_{m=1}^{M}\sum_{j=1}^{N_o}\sum_{k=1}^{N_a} \mathbf{1}_{(H=h_i)} \cdot \mathbf{1}_{(O^m=o_j)} \cdot \mathbf{1}_{(A=a_k)} \cdot \zeta_{i,j,k}\end{aligned}$$

where N_h is the the total number of atomic poses (see Sec.2.2) and h_i is the i-th atomic pose in $\mathcal{H}$ (similarly for N_o, o_j, N_a, and a_k). $\zeta_{i,j,k}$ represents the strength of the interaction between h_i, o_j, and a_k. M is the number of object bounding boxes within the image, and O^m is the object class label of the m-th box.

Modeling Actions. $\phi_2(A,I)$ is parameterized by training an action classifier based on the extended im-

[1] Small distance indicates large image similarity.

age regions of the humans. We have

$$\phi_2(A, I) = \sum_{k=1}^{N_a} \mathbf{1}_{(A=a_k)} \cdot \eta_k^T \cdot s(I) \quad (3)$$

where $s(I)$ is an N_a-dimensional output of a one-vs-all discriminative classifier. η_k is the feature weight corresponding to a_k.

Modeling objects. Inspired by Desai et al. (2009), we model objects in the image using object detection scores in each detection bounding box and the spatial relationships between these boxes. Denoting the detection scores of all the objects for the m-th box as $g(O^m)$, $\phi_3(O, I)$ is parameterized as

$$\phi_3(O, I) = \sum_{m=1}^{M} \sum_{j=1}^{N_o} \mathbf{1}_{(O^m=o_j)} \cdot \gamma_j^T \cdot g(O^m) + \quad (4)$$
$$\sum_{m=1}^{M} \sum_{m'=1}^{M} \sum_{j=1}^{N_o} \sum_{j'=1}^{N_o} \mathbf{1}_{(O^m=o_j)} \cdot \mathbf{1}_{(O^{m'}=o_{j'})} \cdot \gamma_{j,j'}^T \cdot b(O^m, O^{m'})$$

where γ_j is the weights for object o_j. $\gamma_{j,j'}$ encodes the weights for geometric configurations between o_j and $o_{j'}$. $b(O^m, O^{m'})$ is a bin function with a grid representation as in Desai et al. (2009) that models the relationship between the m-th and m'-th bounding boxes.

Modeling human poses. $\phi_4(H, I)$ models the atomic pose that H belongs to and the likelihood of observing image I given that atomic pose. We have

$$\phi_4(H, I) \quad (5)$$
$$= \sum_{i=1}^{N_h} \sum_{l=1}^{L} \mathbf{1}_{(H=h_i)} \cdot \left(\alpha_{i,l}^T \cdot p(\mathbf{x}_I^l | \mathbf{x}_{h_i}^l) + \beta_{i,l}^T \cdot f^l(I)\right)$$

where $\alpha_{i,l}$ and $\beta_{i,l}$ are the weights for the location and appearance of the l-th body part in atomic pose h_i. $p(\mathbf{x}_I^l | \mathbf{x}_{h_i}^l)$ is the Gaussian likelihood of observing $\mathbf{x}_I^l$, the joint of the l-th body part in image I, given the standard joint location of the l-th body part in atomic pose h_i. $f^l(I)$ is the output of a detector for the l-th body part in this image.

Spatial relationship between objects and body parts. We achieve a better modeling of objects and human body parts by considering their spatial relationships. $\phi_5(H, O)$ is parameterized as

$$\phi_5(H, O) \quad (6)$$
$$= \sum_{m=1}^{M} \sum_{i=1}^{N_h} \sum_{j=1}^{N_o} \sum_{l=1}^{L} \mathbf{1}_{(H=h_i)} \cdot \mathbf{1}_{(O^m=o_j)} \cdot \lambda_{i,j,l}^T \cdot b(\mathbf{x}_I^l, O^m)$$

where $b(\mathbf{x}_I^l, O^m)$ denotes the spatial relationship between the l-th body part in I and the m-th object bounding box. We again use the bin function as in Desai et al. (2009). $\lambda_{i,j,l}$ encodes the weights for this relationship when the object class of O^m is o_j.

2.2. Obtaining Atomic Poses

In this section, we discuss a clustering method to obtain atomic poses. Given the training images, we first align the annotations of each image so that the torsos of all the humans have the same position and size, and normalize the range of variations of both position and orientation to $[-1, 1]$. If there is a missing body part due to occlusion, we fill in the annotation with the average annotation values for that particular part. We then use hierarchical clustering with the max linkage measure to obtain a set of clusters, where each cluster represents an atomic pose. Given two images i and j, their distance is measured by $\sum_{l=1}^{L} \mathbf{w}^T \cdot |\mathbf{x}_i^l - \mathbf{x}_j^l|$, where $\mathbf{x}_i^l$ denotes the position and orientation of the l-th body part in image i, $\mathbf{w}$ is a weight vector (0.15 and 0.1 for location and orientation components respectively), L is the number of body parts.

The atomic poses can be thought of as a dictionary of human poses, where the layouts of body parts described by the same atomic pose are similar. Intuitively, human pose estimation performance can be improved by using a prior which is learned from the images of the same atomic pose, as compared to relying on a single model for all the images. Therefore, we estimate the spatial relationship between body parts in the pictorial structure (Felzenszwalb & Huttenlocher, 2005) model for each atomic pose respectively, which will be used in our model inference stage (Sec.2.4).

2.3. Model Learning

Our model (Eqn.1) is a standard Conditional Random Field (CRF) with no hidden variables. We use a belief propagation method (Pearl, 1988) with Gaussian priors to learn the model parameters $\{\zeta, \eta, \gamma, \alpha, \beta, \lambda\}$. All object detectors and body part detectors are trained using the deformable parts model (Felzenszwalb et al., 2010), while the action classifier is trained using the spatial pyramid method (Lazebnik et al., 2006). A constant 1 is appended to each feature vector so that the model can learn biases between different classes.

Conditioned on the image appearance information in $\phi_2 \sim \phi_5$, our model learns the strength of the compatibility between a set of actions, objects, and human poses in ϕ_1. Fig.3 visualizes the connectivity structure learned from the sports dataset (described in Sec.3.1). Each connection is obtained by marginalizing ζ in Eqn.2 with respect to the other concept, e.g.

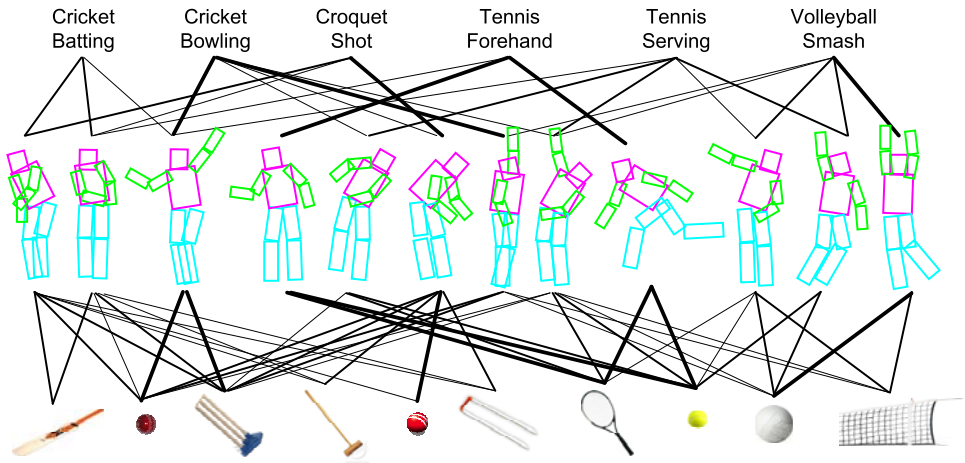


Figure 3. The learned connectivity map of actions, poses, and objects using the sports (Gupta et al., 2009) dataset. Thicker lines indicate stronger connections while thinner connections indicate weaker connections. We did not show the connections between actions and objects because they are tricky (e.g. "tennis serving" connects with "tennis ball" and "tennis racket"). We also ignore connections that are very weak.

the strength of the connection between pose h_i and object o_j is estimated by $\sum_{k=1}^{N_a} \exp(\zeta_{i,j,k})$.

Fig.3 shows that our method learns meaningful action-object-pose interactions, such as the connection between "tennis forehand" and the fourth atomic pose which is a reasonable gesture for the action, the object "volleyball" and the last atomic pose, etc.

2.4. Recognition I: Action Classification, Object Detection, and Pose Estimation

Given a new image, inference on Eqn.1 gives us the results of action classification, object detection, and human pose estimation. We initialize the model inference with the SVM action classification results using the spatial pyramid representation (Lazebnik et al., 2006), object bounding boxes obtained from independent object detectors (Felzenszwalb et al., 2010), as well as initial pose estimation results from a pictorial structure model (Sapp et al., 2010) estimated from all training images, regardless of the belongingness of different atomic poses. We then iteratively perform the following three steps until a local maximum of $\Psi(V, O, H, I)$ is reached.

Updating the layout of human body parts. From the current inference result, we compute the marginal distribution of the human pose over all atomic poses: $\{p_{(H=h_i)}\}_{i=1}^{N_h}$. From this distribution, we refine the prior of the joint location of each body part l in this image using a mixture of Gaussians $\sum_{i=1}^{N_h}[p_{(H=h_i)} \cdot \mathcal{N}(\mathbf{x}_{h_i}^l)]$, where $\mathcal{N}(\mathbf{x}_{h_i}^l)$ is the prior distribution for body part l in the i-th atomic pose estimated in Sec.2.2. Furthermore because the pictorial structure inference can be very efficient if the part dependencies are Gaussians, we use a Gaussian distribution to approximate each mixture of Gaussians. Then we use pictorial structure with these new Gaussian distributions to update the pose estimation results.

Updating the object detections. With the current pose estimation result as well as the marginal distribution of atomic poses and action classes, we use a greedy forward search method (Desai et al., 2009) to update the object detection results. We use (m, j) to denote the score of assigning the m-th object bounding box to object o^j, which is initialized as

$$(m,j) = \sum_{i=1}^{N_h}\sum_{l=1}^{L} p_{(H=h_i)} \cdot \lambda_{i,j,l}^T \cdot b\left(\mathbf{x}_H^l, O^m\right) + \sum_{i=1}^{N_h}\sum_{k=1}^{N_a} p_{(H=h_i)} \cdot p_{(A=a_k)} \cdot \zeta_{i,j,k} + \gamma_j^T \cdot g(O^m) \quad (7)$$

Initializing the labels of all the windows to be background, the forward search repeats the following steps

1. Select $(m^*, j^*) = \arg\max\{(m, j)\}$.
2. Label the m^*-th object detection window as o_{j^*} and remove it from the set of detection windows.
3. Update $(m, j) = (m, j) + \gamma_{j,j^*}^T \cdot b(O^m, O^{m^*}) + \gamma_{j^*,j}^T \cdot b(O^{m^*}, O^m)$.

until $(m^*, j^*) < 0$. After this step, all object bounding boxes are assigned to either an object label or the background.

Updating the action and atomic pose labels. Based on the current pose estimation and object detection results, we optimize $\Psi(A, O, H, I)$ by enumerating all possible combinations of A and H labels.

2.5. Recognition II: Computing Action Distance

Based on our model inference results, we measure the distance between two action images considering not

only action classes but also objects and human poses in the action. For an image I, we use our mutual context model to infer marginal distributions on the action classes $p(A|I)$ and atomic poses $p(H|I)$ respectively. We also obtain a N_o-dimensional vector whose j-th component is set to 1 if the object o_j is detected in image I, or 0 otherwise. We normalize this vector to obtain a distribution $p(O|I)$ for all the objects in this image. We then measure the distance between two images I and I' by

$$2 \cdot D(p(A|I), p(A|I')) + D(p(O|I), p(O|I')) + 2 \cdot D(p(H|I), p(H|I')) \quad (8)$$

where D (described below) indicates the distance between two probability distributions. We assign a lower weight to objects because the performance of object detection is not as good as action classification and pose estimation (Sec.3.2). In this paper we consider two distance measures (D) for probabilities:

- Total variance $T(\mathbf{p}, \mathbf{q}) = \sum_i |p_i - q_i|$.
- Chi square statistic $\chi^2(\mathbf{p}, \mathbf{q}) = \sum_i \frac{(p_i - q_i)^2}{p_i + q_i}$.

Note that our model (Sec.2.1) jointly considers human actions, objects, and human poses, and therefore the probability distribution estimated from each of them considers image appearance as well as contextual information from the other two. Our distance measure further takes into account the three components together. In Sec.3.3.2 we show that our approach captures much semantic level differences between images of human actions and the results are largely consistent with human perceptions as shown in Fig.2.

3. Experiment

3.1. The Six-Class Sports Dataset

We carry out experiments on the six-class sports dataset (Gupta et al., 2009). For each action there are 30 training images and 20 testing images. The objects that we consider are: cricket bat, ball, and stump in "cricket batting" and "cricket bowling"; croquet mallet, ball, and hoop in "croquet shot"; tennis racket and ball in "tennis forehand" and "tennis serving"; volleyball and net in "volleyball smash".

We train an upper-body detector on this dataset using Felzenszwalb et al. (2010). The detector works almost perfectly because of the relatively clean image background. We normalize the images based on the size of the detection boxes such that we do not need to search over scales in human pose estimation. We obtain 12 atomic poses on this dataset (shown in Fig.3).

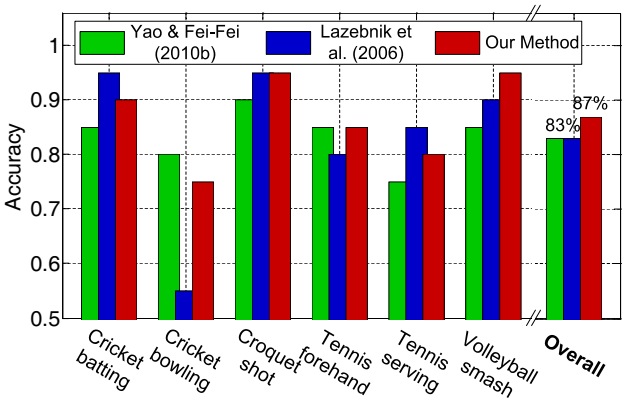


Figure 4. Action classification performance of different methods on the sports dataset.

3.2. Action Classification, Object Detection, and Human Pose Estimation

The action classification results are shown in Fig.4. We also compare our method with other approaches for object detection and human pose estimation in Tbl.1. Following the convention in Ferrari et al. (2008), a body part is considered correctly localized if the endpoints of its segment lie within 50% of the ground-truth segment length from their true positions. As in PASCAL VOC (Everingham et al., 2007), an object detection bounding box is considered correct if the ratio between its intersection with the ground truth and its union with the ground truth is greater than 50%.

We observe that our method achieves better performance than the baselines in almost all experiments. We obtain better action classification and pose estimation results compared to Yao & Fei-Fei (2010b) because we use stronger body part detectors and incorporate the discriminative action classification component in the model of this paper. Please refer to Yao & Fei-Fei (2010b) for more analysis and comparison of the mutual context model and the other approaches.

3.3. Distance between Action Images

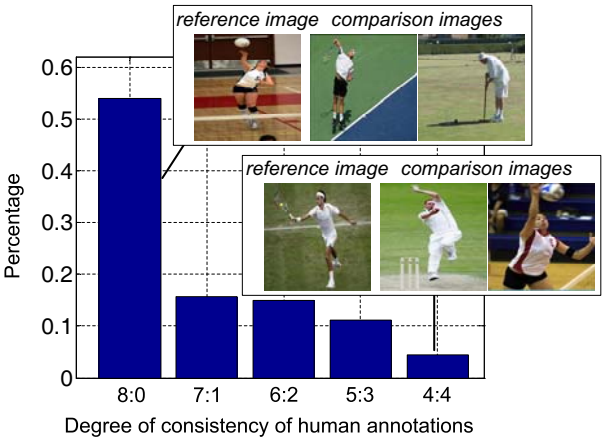
3.3.1. Human Perception of Action Distances

Before we evaluate our distance metric described in Sec.2.5, we study how humans measure the similarity between action images. First, we are interested in whether humans agree with one another on this task. In every trial of our experimental study, we give a human subject one reference image and two comparison

Table 1. Object detection and pose estimation results on the sports data. We use average precision and detection accuracy to measure the performance of object detection and pose estimation respectively. We bold the best performance in each experiment. Note that the object detection result is not directly comparable to that of Yao & Fei-Fei (2010b), because in this paper we detect each object in all testing images, while in that paper the object is only detected in images of the action classes that could contain the object (e.g. detecting "volleyball" in "volleyball smash" images).

Object Detection			
Method	Felzenszwalb et al. (2010)	Desai et al. (2009)	Our Method
cricket bat	17%	18%	**20%**
cricket ball	24%	27%	**32%**
cricket stump	77%	**78%**	77%
croquet mallet	29%	32%	**34%**
croquet ball	50%	52%	**58%**
croquet hoop	15%	17%	**22%**
tennis racket	33%	31%	**37%**
tennis ball	42%	46%	**49%**
volleyball	64%	65%	**67%**
volleyball net	4%	6%	**9%**
overall	36%	37%	**41%**

Human Pose Estimation			
Method	Yao & Fei-Fei (2010b)	Andriluka et al. (2009)	Our Method
head	58%	71%	**76%**
torso	66%	69%	**77%**
left/right upper arms	44%	44%	**52%**
	40%	40%	**45%**
left/right lower arms	27%	35%	**39%**
	29%	36%	**37%**
left/right upper legs	43%	58%	**63%**
	39%	**63%**	61%
left/right lower legs	44%	59%	**60%**
	34%	71%	**77%**
overall	42%	55%	**59%**



(a) X-axis represents the degree of consistency when humans measure the similarity between different action images, e.g. "7:1" means seven of the eight subjects have the same annotation in a given trial. Y-axis is the percentage of the corresponding trials in all the 252 trials. We also show the images in two trials where the degree of consistency is "8:0 (the left comparison image is more similar to the reference image)" and "4:4" respectively.

(b) Examples of action similarities obtained from human annotation. In each row, the reference image is indicated by a yellow bounding box. The magenta numbers are the similarity with the corresponding reference image.

Figure 5. Human annotations of action distance.

images (as shown in Fig.5(a)), and ask the subject to annotate which of the two comparison images is more similar to the reference image. We generate two trials of experiments for every possible combination of action classes from the sports dataset, and therefore our experiment consists of $2 \times (6 + C_{6,2}) = 252$ trials. We give the same 252 trials to eight subjects.

Fig.5(a) summarizes the consistency of the received responses. We observe that in most situations the eight subjects agree with each other (54% 8:0 as compared to 4% 4:4), even in many confusing trials. For example as shown in Fig.5(a), all eight subjects believe the "volleyball smash" image is closer to the "tennis forehand" image than the "croquet shot" image because the former two images have similar human poses.

Having shown that humans tend to give similar annotations in measuring the similarity between different action images, we obtain the ground truth similarity between action images by averaging annotations from different human subjects. We give each subject an annotation task where an image I^{ref} is treated as the reference image for 50 trials. In each trial we randomly select two different test images to compare with I^{ref}. Five of the eight subjects are assigned this task, resulting in 250 pairwise rankings of the 120 test images for

I^{ref}. We then use the edge flow method (Jiang et al., 2011) to convert these pairwise rankings to a similarity vector $\mathbf{s} = \{s(I^{ref}, I^i)\}_{i=1}^{120}$, where $s(I^{ref}, I^i)$ denotes the ground truth similarity between I^{ref} and I^i. We obtain $\mathbf{s}$ by solving an optimization problem

$$\begin{aligned} &\text{minimize} \;\; \mathbf{M} \cdot \mathbf{s} = \mathbf{1} \\ &s.t. \;\; \mathbf{s} \succeq 0, \;\; \|\mathbf{s}\|_2 \leq 1 \end{aligned} \tag{9}$$

where $\mathbf{M}$ is a 250×120 sparse matrix where $M_{j,k} = 1$ and $M_{j,l} = -1$ if the j-th pairwise ranking indicates that I^k is more similar to I^{ref} than I^l.

We repeat the above procedure to obtain a similarity vector for each test image. Fig.5(b) shows examples of action similarities. Note that $s(I^{ref}, I^i)$ is asymmetric because we obtain the similarity values by treating each test image as the reference image separately.

3.3.2. Evaluating the Distance Metric

In this section, we evaluate the approaches of computing the distance between different action images. With the ground truth similarities of each reference image against all the other images obtained from human annotation (Sec.3.3.1), our goal is to automatically find the images that correspond to large similarity (small distance) values.

Our distance metric is evaluated in the following way. Denote the ground truth similarity between an image I and the reference image I^{ref} as $s(I^{ref}, I)$. We have a ground truth ranking of all the images $\{I^{gt_1}, I^{gt_2}, \cdots\}$ such that $s(I^{ref}, I^{gt_i}) \geq s(I^{ref}, I^{gt_j})$ if $i \leq j$. Using our distance metric we obtain another ranking of all the images $\{I^{re_1}, I^{re_2}, \cdots\}$ by sorting their distance with the reference image in ascending order. The precision of using this distance metric to find n neighboring images for I^{ref} is evaluated by $\sum_{i=1}^{n} s(I^{ref}, I^{re_i}) / \sum_{i=1}^{n} s(I^{ref}, I^{gt_i})$. Average precision of using all the test images as reference images is adopted for performance evaluation.

We compare our distance metric (Eqn.8) with a baseline approach that is based on spatial pyramid image classification (Lazebnik et al., 2006). In that approach, an image is represented by the six-dimensional confidence scores obtained from one-vs-all SVM classification. The distance between the confidence scores is used to measure the distance between two images. We also compare our method with some control approaches that use each of the three components (action, object, and human pose) of Eqn.8 individually.

We observe from Fig.6 that our method outperforms the baseline and all the other control settings. The two probability distance measures, χ^2 and T, achieve very similar performance in all the methods. Among the three components, using actions only performs the best while using objects only performs the worst. The reason might be that, objects are usually small such that the human annotations put less weights to objects compared with that of actions or human poses. Also, Tbl.1 shows that object detection does not perform as well as pose estimation or action classification, making it less reliable when using objects only for distance computation.

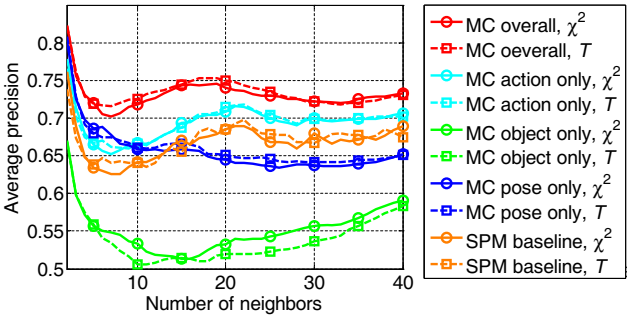


Figure 6. Comparison of different distance metrics evaluated by average precision with respect to the number of similar images in top of the ranking. "MC" denotes "mutual context" and "SPM" is "spatial pyramid matching".

Fig.7 shows the top 20 images obtained using our method and the baseline spatial pyramid method. We observe that our results are significantly more consistent with human perception. Our method can not only find the images that have the same action as the reference image, but also capture the detailed similarity of semantic meaning such as human pose. For example, in Fig.7(b), the "volleyball smash" image returns 17 images of the same action, and the humans in the next 3 images have similar poses as the human in the reference image.

4. Conclusion

In this paper, we show that the joint modeling of actions, objects, and human poses can not only improve the performance of action classification, object detection, and pose estimation, but also lead to an action distance measure approach whose output is largely consistent with human annotations. Our future work will be applying our method on larger datasets.

Acknowledgments

L.F-F. is partially supported by an NSF CAREER grant (IIS-0845230), an ONR MURI grant, the DARPA VIRAT program and the DARPA Mind's Eye

program. B.Y. is partially supported by the SAP Stanford Graduate Fellowship. We also would like to thank Jia Deng and Jia Li for their comments on the paper.

References

Andriluka, M., Roth, S., and Schiele, B. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

Bourdev, L. and Malik, J. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

Delaitre, V., Laptev, I., and Sivic, J. Recognizing human actions in still images: A study of bag-of-features and part-based representations. In *BMVC*, 2010.

Desai, C., Ramanan, D., and Fowlkes, C. Discriminative models for multi-class object layout. In *ICCV*, 2009.

Everingham, M., Gool, L. Van, Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

Felzenszwalb, P. F. and Huttenlocher, D. P. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61 (1):55–79, 2005.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE T. Pattern Anal.*, 32 (9):1627–1645, 2010.

Ferrari, V., Marín-Jiménez, M., and Zisserman, A. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

Gupta, A., Kembhavi, A., and Davis, L. S. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE T. Pattern Anal.*, 31(10):1775–1789, 2009.

Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. Statistical ranking and combinatorial hodge theory. *Math. Program., Ser. B*, 127:203–244, 2011.

Laptev, I. and Mori, G. Statistical and structural recognition of human actions. Tutorial on Human Action Recognition at ECCV, 2010.

Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (2nd ed.)*. Morgan Kaufmann, 1988.

Sapp, B., Jordan, C., and Taskar, B. Adaptive pose priors for pictorial structures. In *CVPR*, 2010.

Yao, B. and Fei-Fei, L. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010a.

Yao, B. and Fei-Fei, L. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010b.

(a) Comparison of our distance metric and the baseline on a "tennis serving" image.

(b) Comparison of our distance metric and the baseline on a "volleyball smash" image.

Figure 7. Examples of the top-ranked images obtained from our method and the baseline approach that only relies on action classification (Lazebnik et al., 2006). In each case, the top-left image surrounded by a yellow rectangle is the reference image, and all the other images are organized in a row major order with respect to ascending distance values to the corresponding reference image.