Recognising Human-Object Interaction via Exemplar based Modelling

Jian-Fang Hu[†], Wei-Shi Zheng[‡*], Jianhuang Lai[‡], Shaogang Gong[◇], and Tao Xiang[◇]
[†]School of Mathematics and Computational Science, Sun Yat-sen University, China
[‡]School of Information Science and Technology, Sun Yat-sen University, China
[*]Guangdong Province Key Laboratory of Computational Science, Guangzhou, China
[◇]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
hujianf@mail2.sysu.edu.cn, wszheng@ieee.org, stsljh@mail.sysu.edu.cn, {sgg,txiang}@eecs.qmul.ac.uk

Abstract

Human action can be recognised from a single still image by modelling Human-object interaction (HOI), which infers the mutual spatial structure information between human and object as well as their appearance. Existing approaches rely heavily on accurate detection of human and object, and estimation of human pose. They are thus sensitive to large variations of human poses, occlusion and unsatisfactory detection of small size objects. To overcome this limitation, a novel exemplar based approach is proposed in this work. Our approach learns a set of spatial pose-object interaction exemplars, which are density functions describing how a person is interacting with a manipulated object for different activities spatially in a probabilistic way. A representation based on our HOI exemplar thus has great potential for being robust to the errors in human/object detection and pose estimation. A new framework consists of a proposed exemplar based HOI descriptor and an activity specific matching model that learns the parameters is formulated for robust human activity recognition. Experiments on two benchmark activity datasets demonstrate that the proposed approach obtains state-of-the-art performance.

1. Introduction

Recently the problem of recognising human action from a single image has received increasing interest [23, 1, 5, 21]. In this context, action can be defined as the Human-Object Interaction (HOI). Existing approaches focus on modelling the co-occurrence or spatial relationship between human and the manipulated object. The co-occurrence relationship, for example, can be modelled by a mutual context model that joins object detection and human pose estimation (i.e. the posture information) together [23]; whilst the spatial relationship concerns more about the relative geometric information, e.g. the relative position and overlap between a human and objects that join human detection or annotation and object detection together [13, 1, 18, 4]. In addition, global context that describes holistic semantic information where HOI takes place in an image is exploited to assist in HOI modelling in most existing works [21, 24, 3, 19]. Beyond still images, there are other works that exploit HOI modeling in the domain of video [7, 17, 11] by incorporating the motion cues for the task. In particular, [11] presents a method for categorising manipulated objects and tracking 3D articulated hand pose in context of each other in order to figure out the interactions between human and interacting objects. In addition to explicitly model the spatial relationship between human and object, the relative motion of object w.r.t human is also exploited to describe their interactions in [16].

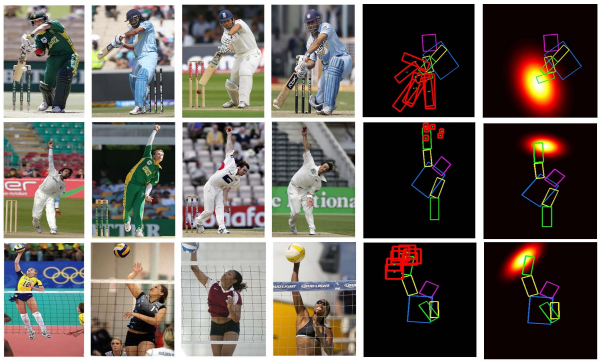

Figure 1. Illustration of spatial pose-object interaction exemplar computation. Each row shows example of an exemplar. **Columns** 1-4 show four images represented by the same atomic pose. **Column** 5 shows manipulated objects locations overlapped with corresponding atomic pose. Red boxes indicate objects. **Column** 6 shows the exemplars. Warmer colors indicate larger response.

*corresponding author

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.390

However, most of the existing HOI modelling approaches rely heavily on explicit human pose estimation [23] or directly using locations of human and objects as HOI representation [13, 1, 18]. Specifically, for the methods that represent action using the spatial relationship between human and object, person and object detections are critical [13, 1, 18]; whilst for those based on the co-occurrence modelling, accurate human pose estimation is crucial [23]. Nevertheless, the problem of detecting objects, especially those small-size objects such as badminton and tennis ball is far from being solved; the problem of estimating human pose under occlusion and large pose variations also remains unsolved. Therefore, the performance of existing approaches is hindered by their HOI representation directly based on human/object detection and pose estimation.

In this paper, we overcome this limitation by proposing a model for learning a set of exemplars to representing human object interaction. Exploring spatial pose-object interaction exemplar is motivated by the observation that for a human activity of similar human poses, the manipulated objects, if there is any, would appear at similar relative positions, i.e. relative to a reference point, such as torso centre of human (see examples in column 5 of Fig. 1). Therefore, the configuration of pose and object can be viewed as an exemplar for describing the action where interaction between human and object happens. This type of exemplars is termed as *spatial pose-object interaction exemplar*.

A spatial pose-object interaction exemplar is mainly represented as a density function that tells how likely an object appears with respect to an (atomic) pose at a position around a person. Some examples of spatial pose-object interaction exemplars can be found in the 4^{th} column of Fig. 1 and Fig. 2. By representing HOI as a set of exemplars, the HOI in an image can be represented by measuring the response of different exemplars within image. Due to the probabilistic modelling for the mutual spatial structure information between human and object in our exemplars, one no longer requires accurate detection of the human and object, and the estimation of human pose. Furthermore, we develop a new activity specific ranking method for recognition. Together with the exemplar based HOI descriptor, this provides a robust still-image based human action recognition framework.

Despite that exemplar based modelling has been applied to a variety of visual recognition problems including scene recognition [10], object detection [14], pose estimation [15], exemplar based HOI modelling has been mostly unexploited. The use of exemplar in existing work is focused on transferring useful information extracted from meta-data to a new data point. This is very different from our objective, which is to develop an exemplar based representation. More recently, an exemplar approach was exploited for action recognition [22]. However, the purpose of exemplar in [22] is for selecting a set of representative samples for each class, which differs from our notion and design of the exemplar in this work. Moreover, compared to [22], our exemplar modelling is to model the mutual structure between a human and an object probabilistically, and crucially our approach does not rely on any feature point annotation/detection and depth information estimation, therefore much more useful and generic for wider scenarios.

We evaluate the effectiveness of our approach on two benchmark datasets: a sports dataset [9] and a people-playing-musical-instrument (PPMI) dataset [21]. Our results show that the proposed approach is able to produce state-of-the-art performance, comparing with most recently proposed competitors. We also demonstrate the robustness of our approach in Sec. 3.5.

2. Approach

Our exemplar modelling consists of two parts: 1) a new exemplar based HOI descriptor (Sec.2.1∼Sec.2.4); and 2) a matching model for learning combination weights of all cues in the proposed HOI descriptor (Sec. 2.5).

2.1. Learning Atomic Poses

Instead of explicit human pose estimation, our modelling is based on the use of a set of atomic poses [23] learned from training data. Atomic poses are representative poses that often occur in specific HOI activities. We assume that each pose involved in the activities can be associated to a most similar atomic pose.

Given a set of M training samples $\mathcal{Q} = \{Q_1, Q_2, \cdots, Q_M\}$ from Z activity classes, we learn the atomic poses by following [23]. The atomic poses are generated based on the annotations of human body parts during training. To derive the atomic poses from annotated training data, we first align all the annotations so that the torsos in all the images have the same position, width and height. Then all the aligned annotations are clustered by Affinity Propagation (AP) clustering method [8]. The computed cluster centres $H = \{H_1, H_2, ..., H_N\}$ form our dictionary of atomic poses, that is, each cluster represents an atomic pose. Some examples of atomic poses we derive from a sports dataset are illustrated in Fig. 1 and Fig. 2. The advantage of using the AP method is that we do not need prior knowledge on the number of atomic poses N, which is determined automatically.

2.2. Constructing Exemplar Dictionary

Given atomic poses, we would like to build a spatial pose-object interaction exemplar dictionary that both encodes and interprets interactions between human and objects. Our idea of exploring interaction exemplar is inspired by the observation that the locations of the manipulated objects are constrained by person's location, pose and type of activity. For example, if a man is playing volleyball as il-

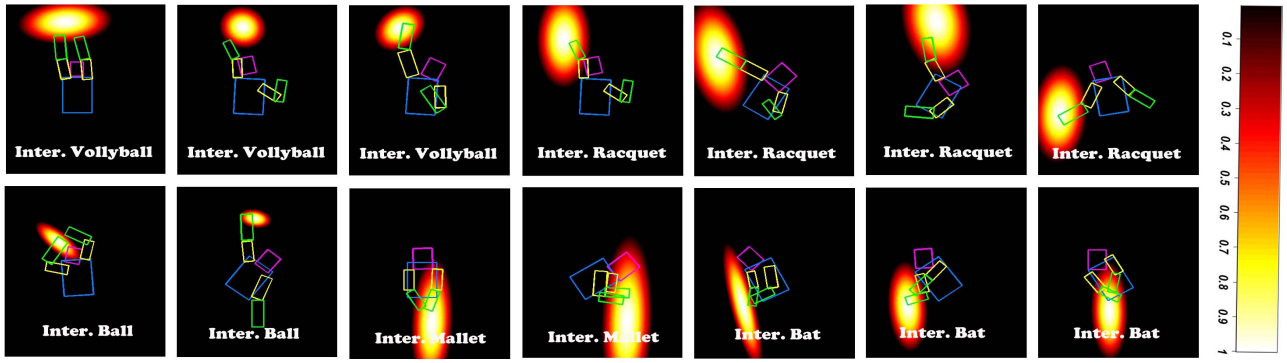

Figure 2. Some examples of spatial pose-object interaction exemplar. All annotated boxes in an image constitute an atomic pose, where different parts are discovered and marked with different colours. The lighting area indicates the distribution of manipulated object.

lustrated in the first picture of Fig. 2, it is more likely that the volleyball would appear near his hands (i.e. the bright region) rather than near his torso or feet. Hence, we formulate a distribution function $G(\mathbf{x})$ to describe the likelihood that a manipulated object would appear at location $\mathbf{x}$ around a person for a specific spatial pose-object interaction. In this work, we call such a distribution as *Exemplar*. By utilising the distribution modelling, we are able to describe the interaction between pose and object in a probabilistic way, rather than directly using the label information or precise coordinates of object and person as features for inference.

We will compute exemplar for each pair of manipulated object and atomic pose appears in the training set. The obtained exemplars are called *spatial exemplar dictionary*. For the N atomic poses and K objects, we can construct a dictionary of spatial pose-object interaction exemplars G_{nk} for all atomic poses H and manipulated objects $O = \{O_k\}_{k=1,2...K}$. We denote it as $D = \{G_{nk}\}_{n=1,2,...,N,k=1,2,,..,K}$.

2.2.1 Dictionary Estimation

We assume the distribution of each elementary exemplar follows normal distribution with parameters $\boldsymbol{\mu}$ and Σ, which are mean vector and covariance matrix, respectively. It is based on the assumption that for each exemplar, object would appear in a similar location relative to a human in an activity, and thus multiple exemplars can be viewed as multi-gaussian distribution for describing the location variation. That is we can formulate density function for an elementary exemplar by

$$G(\mathbf{x}) \propto exp[-(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})] \quad (1)$$

For each training sample $Q \in \mathcal{Q}$, we denote its corresponding atomic pose as H_n and its manipulated object as O_k. We aim to learn a measure of the spatial pose-object interaction exemplar $G(\mathbf{x})$ that tells how likely O_k will locate at position $\mathbf{x}$. Note that all the human and object configurations given in the training set vary in size and position in different samples, i.e., all these data are given in different coordinate frames for different samples. In order to derive a uniform coordinate frame, we need to normalise human and object configurations, so that their torso centres and widths are fixed as $({x_t}^o, {y_t}^o)$ and ${w_t}^o$ respectively. This is achieved by computing

$$\begin{aligned}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \tilde{\mathbf{W}}, \tilde{\mathbf{H}}) =& ({x_t}^o - scale \cdot \mathbf{X}(2), {y_t}^o - scale \cdot \mathbf{Y}(2), 0, 0) \\ &+ scale \cdot (\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{H})\end{aligned} \quad (2)$$

where $scale = {w_t}^o/\mathbf{W}(2)$, $\mathbf{X}$ and $\mathbf{Y}$ are vectors that indicate the x-axis and y-axis of body parts and object center respectively, $\mathbf{W}$ and $\mathbf{H}$ are body parts and object width and height respectively. $\mathbf{X}(2)$ and $\mathbf{Y}(2)$ are the x-axis and y-axis of torso centre of the corresponding training data, $\mathbf{W}(2)$ indicates width of the torso, $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \tilde{\mathbf{W}}, \tilde{\mathbf{H}})$ is the normalised configuration. We normalise the configurations only using torso width, because samples represented by the same atomic pose would usually have similar relative width-height ratio for each part and object.

Let $\mathcal{Q}_{nk}$ be a subset of training samples from $\mathcal{Q}$, each of them associating to atomic pose H_n and object O_k. Let $N_{nk} = \#\mathcal{Q}_{nk}$. Now we estimate the Gaussian parameters in the spatial pose-object interaction exemplar (Eq. (1)) using maximum likelihood. For convenience, we denote $\tilde{\mathbf{L}}_i$ as the object location of $i-th$ sample in $\mathcal{Q}_{nk}$. Then the estimation of $(\boldsymbol{\mu}_{nk}, \Sigma_{nk})$ is given by

$$\boldsymbol{\mu}_{nk} = N_{nk}^{-1}\sum_{i=1}^{N_{nk}}\tilde{\mathbf{L}}_i, \Sigma_{nk} = N_{nk}^{-1}\sum_{i=1}^{N_{nk}}(\tilde{\mathbf{L}}_i - \boldsymbol{\mu}_{nk})(\tilde{\mathbf{L}}_i - \boldsymbol{\mu}_{nk})^T$$

To make the estimation more robust, a regularised covariance matrix is modelled as follows:

$$\Sigma_{nk} \leftarrow \lambda\Sigma_{nk} + (1-\lambda)diag(MW^2/2, MH^2/2) \quad (3)$$

where we set $\lambda = 2trace(\Sigma_{nk})/(2trace(\Sigma_{nk}) + MW^2 + MH^2)$, MW and MH are average width and height of object configurations respectively.

After determining $(\boldsymbol{\mu}_{nk}, \Sigma_{nk})$ for each pair of atom-

ic pose H_n and object O_k, we can get the corresponding spatial pose-object interaction exemplar and denote it as $G_{nk}(\mathbf{x})$, which can be considered as a measure of probability of object O_k appearing at location $\mathbf{x}$ relative to the torso centre (x_t^0, y_t^0).

Some examples of the learned spatial pose-object interaction exemplars are visualised in Fig. 2. This figure shows that an atomic pose can interact with two objects or even more, and an object can also interact with multiple atomic poses. However, for each pair of pose and manipulated object, there is only one interaction exemplar to describe the interaction between them. In addition, from this figure, we can observe that the spatial pose-object interaction exemplar can capture some semantic information that tells us how the actor is manipulating the object.

2.3. Inferring Spatial Pose-Object Interaction Using Exemplars

After constructing the exemplar dictionary, we can use the learned dictionary to compute a representation for an HOI activity in a probe image. As aforementioned, the exemplar approach is exploited to avoid estimation of human pose in the probe image and thus nominate the most similar pose information that is contained in our spatial exemplar dictionary for the probe HOI. Based on the nominated atomic poses, the model selects the candidate exemplar in the dictionary and computes the response of probe HOI against exemplar. Finally the model forms a code vector for each probe HOI consisting of all the response of all the exemplars in the dictionary. In the following, we detail the whole process which is also illustrated in Fig. 3.

2.3.1 Nominating Similar Atomic Poses

For each probe HOI, we nominate the most similar atomic poses defined in the spatial exemplar dictionary. For each detected person P in the probe HOI, we first score each training image with $Sim(P, P^i)$, where $Sim(P, P^i)$ is a function that measures the pose similarity between P and P^i, where P^i indicates the person of interest in the i^{th} training image. Note that each person in the training image in our dataset is associated to an atomic pose. Hence S exemplars $\{Tr_{i_s}\}$,$s = 1, 2, ...S$ corresponding to the top scores of $\{Sim(P, P^i)\}_{i=1,\cdots,N}$ are selected, where the effect of S will be evaluated and discussed Sec. 3.4. To compute $Sim(P, P^i)$, we compute the inverse of the distance between their feature representations encoded by pyramid histogram of words (PHOW) [2]. For pyramid histogram of words (PHOW), we extract dense SIFT features, learn a vocabulary of size 512, and finally compute the histogram under three pyramid levels. Here, we further expand the PHOW feature involved to a vector of dimension 32256 using an approximated kernel map for the Chi-Square kernel [20]. It is suggested that pyramid image features can capture soft pose information [1]. Here only 6 parts from upper body are considered for learning the atomic poses, since sometimes only upper body is visible for person of interest.

2.3.2 Computing the Exemplar Response

After selecting the S candidate exemplars $\{Tr_{i_s}\}$,$s = 1, 2, ...S$, we are now computing their response for each probe HOI. First, for each probe HOI in an image, a pretrained torso detector is employed to run on each detected person in the image to obtain the predicted torso box (x_t, y_t, w_t, h_t), where x_t and y_t are x-axis and y-axis of human centre respectively, and w_t and h_t indicate the width and height of the torso respectively. Note that all spatial pose-object interaction exemplars are constructed under the hypothesis that the involved torso locates at $({x_t}^0, {y_t}^0)$ with a width of ${w_t}^0$

Second, for the k^{th} object type O_k, we detect this type of object and predict the most likely existing location (x, y) in the image, which corresponds to the largest detection score denoted by $\mathbf{O}(k)$. Hence an object detection vector $\mathbf{O}$ will be formed for a probe image over all object types.

Third, for each object type O_k and the selected atomic pose H_n, we align the exemplar G_{nk} so that its torso position is (x_t, y_t) and the width is w_t computed by

$$\tilde{G}_{nk}(x, y) = G_{nk}(x/scale + {x_t}^0 - x_t, y/scale + {y_t}^0 - y_t) \quad (4)$$

where $scale = w_t/{w_t}^0$. $\tilde{G}_{nk}(x, y)$ provides a measure of the probability of object O_k appearing at (x, y) in the image given atomic pose H_n. Larger value means that O_k would more likely appear at (x, y) (see Fig. 6 column 2 for examples of $\tilde{G}$). After alignment, we update the detected object location (x_o, y_o) with respect to $\tilde{G}_{nk}$ and compute the corresponding semantic spatial interaction response as follows

$$\mathbf{I}(n, k) = \tilde{G}_{nk}(x_o, y_o). \quad (5)$$

We compute Eq. (5) for each selected candidate atomic poses and each object type. Then we can obtain a matrix $\mathbf{I}$ of size $N \times K$. Each entry of this matrix represents the response with respect to the corresponding atomic pose and object category, where entries corresponding to non-selected atomic pose are zero. The obtained matrix is then reshaped as a vector $\mathbf{I}$.

2.4. A HOI Descriptor

The spatial exemplar response vector $\mathbf{I}$ as described in Sec. 2.3.2 can only tell the mutual spatial structure information, i.e. the probabilistic geometric information between human and object. It does not capture information about the pose and the object, which is also important for describing HOI. Hence, in the final HOI descriptor, we include the pose appearance feature $\mathbf{P}$ and object detection vector $\mathbf{O}$. These two information help compute the confidence of human pose profiling and object's existence, re-

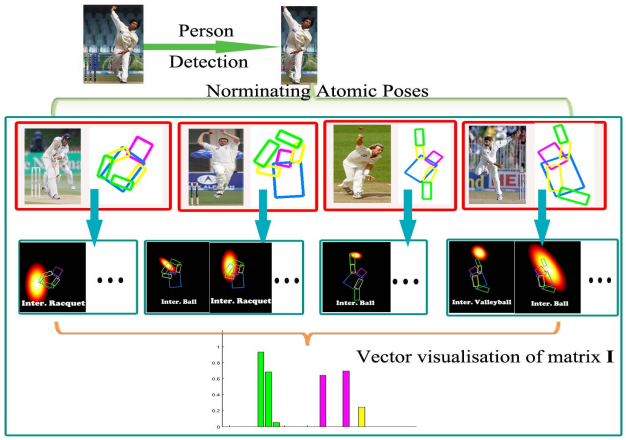

Figure 3. A graphical illustration of computing exemplar response. The last row is a vector visualisation of matrix $\mathbf{I}$ in Eq. (5). For better visualization, bars that associate to different manipulated objects are marked with different colors: cricket bat (red), cricket ball (green), croquet mallet (blue), tennis racket (magenta), volleyball (yellow). From the final representation, we can observe that the actor is manipulating a tennis racket or cricket ball.

spectively to define a HOI. The combination of $\mathbf{I}$, $\mathbf{P}$, and $\mathbf{O}$ are indeed necessary because they provide complementary information to each other, where $\mathbf{I}$ indicates spatial interaction response and $[\mathbf{P}; \mathbf{O}]$ indicates appearance interaction response. Thus, we define this combination as our *full interaction descriptor*.

In addition, similar to existing approach [23, 1, 5, 9], we also combine the contextual features. In summary, our HOI descriptor $\mathbf{H}$ has the following three parts: 1) the spatial pose-object exemplar response vector $\mathbf{I}$ as introduced in the last two sections; 2) the appearance interaction response including pose descriptor $\mathbf{P}$ and the object detection score vectors $\mathbf{O}$; 3) the scene contextual information around a person $\mathbf{C}$. For pose component and scene context, we simply extract pyramid histogram of words (PHOW) from the person of interest and global image respectively. Our final HOI descriptor can be formulated as follows

$$\mathbf{H} = [\mathbf{I}; \mathbf{P}; \mathbf{O}; \mathbf{C}] \tag{6}$$

Compared to existing HOI descriptors [23, 1, 5, 9], the proposed one mainly differs in the use of spatial exemplar response $\mathbf{I}$, while the rest three terms are exploited in existing work in a similar way [1, 5, 9]. Note that not all four parts of the descriptor are equally informative for representing HOI. In the next section, a matching model is proposed which implicitly perform feature selection.

2.5. Matching Model

We wish to quantify all the cues in our HOI descriptor so as to mine as much information as possible for our activity analysis. Now, we have four components for each HOI descriptor. For each component, we learn an one-vs-all discriminative classifier over Z activity classes $\mathcal{C} = \{a_1, a_2, ...a_Z\}$, and hence we would obtain 4 one-vs-all discriminative classifiers. Then, for each activity sample, a 4 dimensional vector denoted as $\mathbf{s}_a$ is formed, which consists of outputs of the 4 classifiers with respect to each class $a \in \mathcal{C}$. Based on them, a weight vector $\mathbf{w}_a$ is formed for each class to combine those outputs of the 4 classifiers and therefore a prediction score $\mathbf{w}_a^T \mathbf{s}_a$ is computed for each class. We assign the class label corresponding to the largest prediction score to a probe as follows:

$$a^* = \arg\max_{a \in \mathcal{C}} \mathbf{w}_a^T \mathbf{s}_a \tag{7}$$

In order to get the best prediction, we wish that the true prediction has higher scores than the incorrect one. We learn the parameters $\mathbf{w}_a$ in a large margin framework with the constraint that the prediction score of incorrect hypothesis is lower than the one of the correct hypothesis by at least 1 minus the loss ξ^i as follows

$$\begin{gathered} \min \frac{1}{2}\sum_{z=1}^{Z} \|\mathbf{w}_{a_z}\|^2 + \frac{1}{vM}\sum_{i=1}^{M}\xi^i, \\ s.t. {\mathbf{w}_{a_i}}^T {\mathbf{s}_{a_i}}^i \geq {\mathbf{w}_a}^T {\mathbf{s}_a}^i + 1 - \xi^i,\ \xi^i \geq 0 \\ \forall i = 1, 2, ...M, a \in \mathcal{C}\ /\{a_i\}, \end{gathered} \tag{8}$$

where a_i is the ground truth label of the i^{th} training sample, M represents the training sample number and s_*^i represents the confidences that classifiers assign the sample to class $*$ and v is a parameter to control the trade-off between training error minimization and margin maximization. We set v to be 0.07 in our experiment.

Solving the above quadratic programming problem directly is not easy. However, inspired by [25], we can utilise the one-class SVM toolbox to compute the solution equivalently by applying a simple transformation. Let $\mathbf{w} = [{\mathbf{w}_{a_1}}^T, {\mathbf{w}_{a_2}}^T, ..., {\mathbf{w}_{a_Z}}^T]^T$, $\phi(a_i) = [\mathbf{0}^T, ..., {\mathbf{s}_{a_i}^i}^T, \mathbf{0}^T, ...\mathbf{0}^T]^T$ and $\mathbf{S}_{aa_i}^i = \phi(a_i) - \phi(a)$, where $\mathbf{0}$ is a zero vector. Then Criterion (8) can be rewritten as

$$\begin{gathered} \min \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{vM}\sum_{i=1}^{M}\xi^i, \\ s.t. \mathbf{w}^T \mathbf{S}_{aa_i}^i \geq 1 - \xi^i, \xi^i \geq 0,\ \forall i = 1, \cdots, M,\ a \in \mathcal{C}\ /\{a_i\}. \end{gathered}$$

Let $\mathbf{S} = \{\mathbf{S}_{aa_i}^i\}_{i=1,2,...M, a \in \mathcal{C}\ /\{a_i\}}$. Note that the solution of above model would linearly separate $\mathbf{S}$ from the origin with maximum margin, so we can solve the problem using any general one-class SVM solver.

3. Experiments

3.1. Settings

Dataset. We evaluate our method on two benchmark data sets of HOI activities: a sports data set [9] and a people-playing-musical-instrument (PPMI) data set [21]. The s-

ports data set consists of 300 images of six HOI activities (tennis-forehand, tennis-serve, volleyball-smash, cricket-bowling, cricket-defensive shot, croquet-shot). We follow the same experiment setting as in [23, 1, 5, 9], where for each activity 30 images were selected for training and 20 were selected for testing. As in [23], only five object classes, cricket bat, bowling ball, croquet mallet, tennis racket, volleyball, were employed to model and evaluate HOI for action recognition. For PPMI, there are twelve musical instruments; each image contains people playing an instrument or holding an instrument. The data set contains 2400 images for training and 2400 images for testing [21]. We follow the setting in [23] to select a subset of the data set, where the person of interest can be detected by people detector for experiment. Therefore, we have 2175 images for training and 2035 images for testing in our case.

Settings. We need to detect human, body parts and objects using the deformable part model [6] for sports dataset and PPMI dataset. To train detectors of human, head, torso and upper body, the ground truth bounding boxes in the sports and PPMI were used to generate positive examples, whilst the negative samples were generated from VOC2012. To facilitate reliable detection of person across a variety of poses, we follow [6, 1] and combine detection windows returned by 4 detectors: head detector, torso detector, upper body detector and people detector. Similar to [6], a linear regression method is employed to predict the final human location. Regarding detection of objects, for each object type, we use the corresponding trained detector to obtain the centre location (x_o, y_o) of the object. In order to rely less on the object detection performance, we only use the detected location to represent object without using its scale at this step.

In addition, the number of candidate exemplars for computing the exemplar response in Sec.2.3.2, namely the parameter S, is set to be 3 for sports dataset and 20 for PPMI, which are almost a quarter of number of the learned atomic poses. Its effect would be further evaluated in Sec. 3.4 .

3.2. Sports data set

Here, we report the recognition results of our method on the sports data set. We also compare our method with the following methods: ***Yao*** [23], ***Prest*** [1], ***Desai*** [5] and ***Gupta*** [9]. All these methods utilise pose, object, relation between pose and object and contextual information. They need to use locations of people and human as features [1, 9] or depend on explicit human pose estimation [23, 5].

Table 1 shows the results. It can be seen that our proposed model achieves the best performance and outperforms the state-of-the-art [23] by 5.5%. In comparison, ours improves by 13.6%, 10% and 9.5% as compared to [9], [5] and [1], respectively. The confusion table of our model is shown in Fig. 4. We can observe that our model achieves perfect results on activities of cricket-batting and croquet. It is noted that serious false detection and occlusion can still affect the performance of our model. For example, for classification of volleyball-smash and cricket-bowling, our model achieves lower performances ($\leq$90%) (see Fig. 4). We believe that this is largely due to faulty object detection and actor prediction. For images of cricket-bowling, it is not easy to detect cricket ball, which is small and sometimes partially occluded by the actor's hand. While for image of volleyball-smash, it is often difficult to correctly locate the person of interest, because the image often contains multiple persons performing different actions. These can limit the performance of our HOI descriptor. Some qualitative results are shown in Fig. 6 which demonstrates that spatial pose-object interaction exemplars are able to effectively describe how a person is interacting with a manipulated object for different activities.

Method	Yao [23]	Desai [5]	Prest [1]	Gupta [9]	Our Model
Accuracy (%)	87	82.5	83	78.9	92.5

Table 1. Comparison on the Sports dataset.

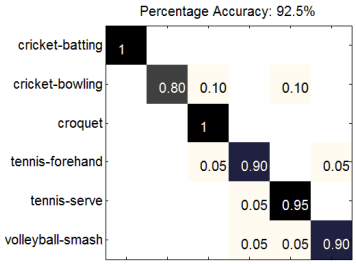

Figure 4. Confusion table of our method on sports dataset.

3.3. PPMI data set

In this experiment, we evaluate different methods on the 24-class classification on PPMI data set. Since the annotations of the dataset used in [23] are not available from the authors, we have to re-annotate this dataset in the same way as what have been done for the sports data set. Specifically, for each training image, we annotated manipulated objects and six body parts, including head, torso, left upper arm, left lower arm, right upper arm and, right lower arm. The best efforts have been taken to follow the exactly the same experimental setting as described in [23].

For comparison, we tabulate the results reported in [23]. Our results are presented in Table 2. From this table, our proposed model achieves 49.34% in average accuracy and 47.56% in mAP (mean average precision) which outperforms SPM [12] and Grouplet [21] by 6% to 8%. The proposed model performs comparably to the state-of-the-art Yao's method [23] on this dataset in terms of mAP. We would like to point out that there could be some bias in such a comparison, because the annotation data were not released and we had to re-do the annotation. Since the confusion

Method	SPM [12, 23]	Grouplet [21, 23]	Yao [23]	Our Model
Accuracy (%)	-	-	-	49.34
mAP (%)	40	42	48	47.56

Table 2. Comparison on the PPMI dataset.

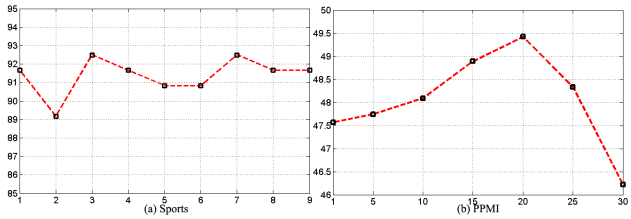

Figure 5. Performance (%) of our models under different numbers of candidate exemplar (Sec. 2.3.1) on Sports and PPMI datasets.

Accuracy (mAP)	Sports	PPMI
Without Perturbation	92.5 (95.25)	49.34 (47.56)
With Perturbation ($\pm$3 pixels)	91.67 (95.74)	48.73 (47.49)
With Perturbation ($\pm$5 pixels)	89.17 (95.42)	48.83 (47.56)
With Perturbation ($\pm$10 pixels)	88.30 (94.93)	48.67 (47.29)

Table 3. Affect of Perturbations (3, 5 and 10 pixels in maximum) to Our Model in Classification (%).

matrix is a 24×24 table, we thus select to present it in a supplementary material due to space constraint.

3.4. Effects of the Number of Candidate Exemplars

We study the effect of different numbers of exemplars S used when nominating atomic poses (Sec. 2.3.1). Fig. 5 (a) and (b) plot the performances of our proposed model on sports and PPMI datasets respectively. The performance reaches the best when $S = 3$ on Sports dataset and when $S = 20$ on the PPMI dataset, which is almost a quarter of the number of atomic poses learned for each dataset and also set as default value in our experiment. Overall, the performance first increases and then decreases as S increases. In comparison, the performance of the model is more sensitive to S on the PPMI dataset. This is because PPMI has more variations of human pose. Due to this fact, a better performance on PPMI is obtained for a larger S, as more candidate exemplars are needed to describe the spatial pose-object interaction in an image.

3.5. Influence of Perturbation in Detections

Here, we evaluate the robustness of our model given errors in person/object detection. In this experiment, a random perturbation ranging from $-p$ to p in pixels is introduced to disturb the relative position between the detected object and human. We test the case when $p = 3$, $p = 5$ and $p = 10$. The results are listed in Table 3, tabulating both accuracy and mAP results where mAP results are in the bracket. The results show that performance drops only slightly by 1%$\sim$4% in accuracy and less than 1% in mAP. Especially, when $p = 3$, there is almost no performance change in mAP given the added detection errors. Note that, even with ± 10 random perturbation in pixels, our model still outperforms the others on Sports dataset (1.3%$\sim$8% more than the compared in accuracy), and outperform SPM and Grouplet on PPMI by 5%$\sim$7% more in mAP and still perform comparably with Yao's method.

3.6. Effect of Exemplar Modelling

We evaluate the effectiveness of exemplar based semantic spatial interaction response by removing the spatial exemplar response vector $\mathbf{I}$ from our HOI descriptor and feeding the rest into our matching model. Unsurprisingly, an accuracy decrease of about $4 \sim 5\%$ of the performance of the full model is observed on the Sports and PPMI datasets. We also test the influence of our full interaction descriptor (combination of spatial interaction response $\mathbf{I}$ and appearance response $[\mathbf{P}; \mathbf{O}]$ as defined in Sec. 2.4) by removing it from our HOI descriptor, and we observe a decrease of about $8 \sim 15\%$ of the performance of the full model. These demonstrate the usefulness of our exemplar modelling.

4. Conclusion and Future work

We have proposed to represent human-object interactions using a set of spatial pose-object interaction exemplars and form a new HOI descriptor, where weight parameters for each component are learned by an activity specific ranking model. A key characteristic of our exemplar based approach is that it models the mutual spatial structure between human and object in a probabilistic way, so as to avert explicit human pose estimation and alleviate the effects of faulty detection of object and human. Our experimental results suggest that our exemplar approach outperforms existing related HOI techniques or perform comparable to them for action recognition from still images. On-going work includes further improvement of the exemplar learning. Specially, our approach depends on the use of atomic poses. However, for some activities, e.g. repairing bike and phoning, it is not easy to mine a set of representative atomic poses from limited data. Hence, in future, we consider exploring the use of large scale data for learning exemplars.

Acknowledgment

We would like to thank Dr. Xiaoming Chen for helpful discussions. This work was supported by the National Natural Science Foundation of China (No. 61102111, 61173084), the 12th Five-year Plan China S&T Supporting Programme (No. 2012BAK16B06), Guangdong Natural Science Foundation(No. S2012010009926), Guangzhou Pearl River Science & Technology Rising Star Project (No. 2013J2200068) and the Guangdong Provincial Government of China through the Computational Science Innovative Research Team programme.

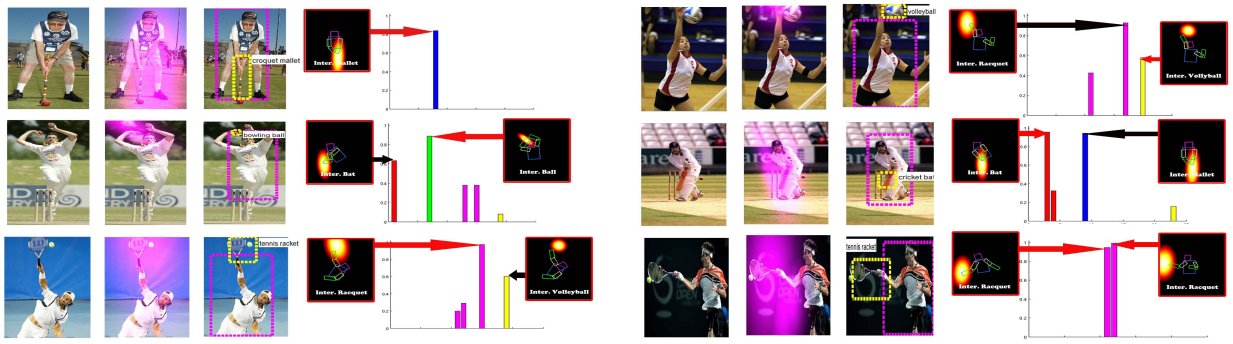

Figure 6. Results for activity interpretation. Each row shows two activity classes. For each class, image of **Column 1** shows HOI activity, image of **Column 2** shows visual response to a normalised pose-object exemplar ($\tilde{G}_{nk}$ in Eq.(4)), image of **Column 3** shows the manipulated object (what) and person (who), image of **Column 4** is a histogram visual result of the pose-object spatial interaction response ($\mathbf{I}$ in Eq.(6), reshaped as a vector). The X-axis and Y-axis of histogram figure are pose-object spatial exemplar index and response value respectively. Exemplars with large response value (>0.5) are presented beside the bar graph. Bars that represent different objects are marked with different colours: cricket bat (red), cricket ball (green), croquet mallet (blue), tennis racket (magenta), volleyball (yellow). Arrows with red colour indicate that the exemplar's manipulated object is consistent with predicted activity type. It illustrates that our exemplar response can provide some semantic information for the activity, which can tell us how the person manipulates the object.

References

[1] C. S. A. Prest and J. Malik. Weakly supervised learning of interactions between humans and objects. *TPAMI*, 34(3):601–614, 2012. 1, 2, 4, 5, 6

[2] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. *ICCV*, 2007. 4

[3] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *Proc. BMVC*, 2010. 1

[4] V. Delaitre, J. Sivic, I. Laptev, et al. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011. 1

[5] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. *In Workshop on Structured Models in Computer Vision*, 2010. 1, 5, 6

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010. 6

[7] R. Filipovych and E. Ribeiro. Recognizing primitive interactions by exploring actor-object states. In *CVPR*, 2008. 1

[8] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007. 2

[9] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *TPAMI*, 31(10):1775–1789, 2009. 2, 5, 6

[10] J. H. J. Xiao, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *CVPR*, 2010. 2

[11] H. Kjellström, J. Romero, and D. Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *CVIU*, 115(1):81–90, 2011. 1

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 6, 7

[13] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. *CVPR*, 2011. 1, 2

[14] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. *ICCV*, 2011. 2

[15] G. Mori and J. Malik. Recovering 3d human body configurations using shape contexts. *TPAMI*, 28(7):1052–1062, 2006. 2

[16] A. Prest, V. Ferrari, and C. Schmid. Explicit modeling of human-object interactions in realistic videos. *TPAMI*, 35(4):835–848, 2013. 1

[17] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012. 1

[18] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 1, 2

[19] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *CVPR*, 2012. 1

[20] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *TPAMI*, 34(3):480–492, 2012. 4

[21] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. *CVPR*, 2010. 1, 2, 5, 6, 7

[22] B. Yao and L. Fei-Fei. Action recognition with exemplar based 2.5 d graph matching. In *ECCV*. 2012. 2

[23] B. Yao and L. Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *TPAMI*, 34(9):1691–1703, 2012. 1, 2, 5, 6, 7

[24] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011. 1

[25] W.-S. Zheng, S. Gong, and T. Xiang. Quantifying and transferring contextual information in object detection. *TPAMI*, 34(4):762–777, 2012. 5